Title

The Algorithmic Management of Polarization and Violence on Social Media

Authors

Stray, Jonathan
Iyer, Ravi
Puig Larrauri, Helena

Publication Date

2023-05-25

Peer reviewed

The Algorithmic Management of Polarization and Violence on Social Media

Jonathan Stray
UC Berkeley Center for Human-compatible AI

Ravi Iyer
University of Southern California Neely Center

Helena Puig Larrauri
Build Up

NOTE: This is a working DRAFT of a paper that has been accepted for publication on the Knight First Amendment Institute at Columbia University website, but which is still in the review and editing process. As such, edits and changes may be made before final publication.

Summary

Social media platforms are involved in all aspects of social life, including in conflict settings. Incidental choices about how they are designed can have profound effects on people when conflict has the potential to escalate to violence. We review theories of conflict escalation and the practice of professional peacebuilders, and distinguish between constructive conflict, which can be part of important societal changes, and destructive conflict where positions become more identity based and intractable. Platforms have largely responded to conflict through content moderation thus far, yet moderation will never affect more than a small amount of objectively policy-violating content, and expanding those efforts will only lead to more backtracking, biased enforcement, and controversy. Instead we draw on recently-published platform experiments, the reports of content creators, international peacebuilding practitioners, and the experiences of those in conflict settings to argue that platforms often incentivize conflict actors toward more divisive and potentially violence-inducing speech, while also facilitating mass harassment and manipulation. We propose that platforms monitor for the conflict relevant side effects of prioritizing distribution based on engagement, such as the incentivization of divisive content, and that they deprecate certain engagement signals (such as comments, shares or time spent) in sensitive contexts. It may also be possible for platforms to support the transformation from destructive to constructive conflict by drawing attention to cross-cutting content, and supporting the on-platform efforts of conflict transformation professionals. To produce widespread legitimacy for these efforts, and overcome the problem of business incentives, we recommend the public creation of clear guidelines for conflict-sensitive platform design, including new kinds of practical conflict metrics.

Introduction

Polarization, violence and social media are inextricably intertwined. Facebook commissioned and agreed with an independent report that concluded that its platform was used to foment division and incite offline violence in Myanmar (Warofka, 2018), and the same military groups that used the platform to foment violence would later restrict it to prevent opposition to a military coup (Wong, 2021). A sitting US President was deplatformed by Twitter with the company acknowledging the use of its platform to incite violence based on fraudulent claims (Twitter, 2021), yet the same platform was credited with being instrumental in protests against less legitimate governments (Tufecki, 2018). Positive or negative, the power of social media to affect conflict is clear.

So far, social media platforms have mostly responded to the problem of violent conflict through content moderation. These efforts are generally reactive, focussing on specific content or crises and outbreaks of violence. Instead, we argue for the prevention of destructive society-scale conflict before escalation to physical violence occurs. Our approach is proactive, long-term, scalable, and operates through platform design rather than content moderation policy. We propose addressing underlying conflict drivers at a deeper level, with an analysis rooted in general conflict principles, informed by the experiences of professional peacebuilders. Peacebuilders are civil society practitioners who use non-violent means to reconcile differences and to collectively transform societal relationships and structures (as distinguished from peacekeeping, which refers to militarized security operations).

Social media companies did not originally envision the central role that their platforms would play in geopolitical and intercommunal conflict and these effects have arisen largely as a result of incidental decisions in service of business goals. However, evidence is accumulating for the nature of the relationship between social media and political conflict. Recent systematic reviews find a positive correlation between social media use and polarization (Kubin *et al.* 2021) but also positive correlations with political knowledge and participation (Lorenz-Spreen *et al.* 2021). Platform experiments in this area are starting to become public, as when Facebook attempted to reduce the distribution of political content (Glazer et. al, 2023; Klepper & Seitz 2021; Gizmodo, 2022). We also have the documented experiences of those living in conflict settings and how they relate to social media (Schirch ed., 2021; Hagey & Horwitz 2021; Build Up 2019; Build Up 2022). This collective evidence has provided an important opportunity to reassess how the design of platforms relates to conflict.

Designing a platform for “better” conflict outcomes requires three things, corresponding to the three sections of this paper.

First, since not all types of conflict are inherently “bad,” we need to be more specific about our design goals. In the first section we review the fundamentals of conflict escalation, showing that large scale changes in perceptions, patterns of behavior, and societal structures occur long before the onset of physical violence. Even if the only goal is to prevent violence, social media must contend with conflict dynamics in much earlier stages of escalation. To clarify what to do in earlier stages we summarize previous discussions of the difference between “constructive” and “destructive” conflict. This includes distinguishing between “affective” polarization, where people dislike and demonize each other, and “issue” based polarization, where people disagree about specific issues. We argue that affective polarization is a reasonable place to start measuring and intervening in pre-violent platform conflict dynamics.

Second, we review the different pathways whereby platform design can facilitate conflict. In our view, there is not strong support for widespread effects from “filter bubbles.” We also don’t think conflict escalation can be addressed through more accurate or more aggressive content moderation, although this may be necessary when conflict is at a violent peak. Instead, we focus on the incentivization of the production and distribution of divisive content, and the ways that platform design can enable mass harassment and manipulation.

Finally, we synthesize the above sections to describe a number of platform design strategies that could result in healthier conflicts, focussing on three types of changes:

- Change content ranking to reward productive and connecting interactions, rather than rewarding divisive content with greater distribution.

- Place reasonable limits on the use of the platform to disseminate broad messages, to better mirror the safeguards of offline life.

- Consider design affordances that support the on-platform work of peacebuilders, recognising that peace is not just the absence of violent conflict, but a society in which everyone can thrive.

In order to design and evaluate effective changes, we will need a new set of conflict-aware metrics to help us understand the incentives and capabilities that platforms create and hold platforms publicly accountable for any resulting externalities. We conclude by discussing other barriers to implementation, and how future research can help.

How Conflicts Escalate

In order to talk about what platforms are and aren’t currently doing to respond to conflict, we need a framework for what conflict is, when it is undesirable, and how it escalates to physical violence. In this section, we draw from the understandings of conflict developed within the professional peacebuilding community, and by researchers in political science and social psychology.

Conflict cycles begin long before violence

In the conflict literature, a number of models look to explain the life-cycle of conflict and its complex dynamics. These models differ in scope and language, but share an important characteristic: that conflict happens in a reinforcing cycle, or spiral. All of them describe the strengthening of factions, hardening of positions, and increasing distrust and fear. If these differences cannot be resolved, conflict participants may resort to violence.

Deutsch (1969) notes that conflict can occur for a variety of reasons, not just incompatibility of goals. Two parties may disagree on the best method to achieve some outcome, or may misperceive each other's true positions, or the true state of the world. Regardless of how a conflict begins, it can take on a life of its own:

> Destructive conflict is characterized by a tendency to expand and to escalate. As a result, such conflict often becomes independent of its initiating causes and is likely to continue after these have become irrelevant or have been forgotten.
>
> …
>
> Paralleling the expansion of the scope of conflict there is an increasing reliance upon a strategy of power and upon the tactics of threat, coercion, and deception. Correspondingly, there is a shift away from a strategy of persuasion and from the tactics of conciliation, minimizing differences, and enhancing mutual understanding and good-will. And within each of the conflicting parties, there is increasing pressure for uniformity of opinion and a tendency for leadership and control to be taken away from those elements that are more conciliatory and invested in those who are militantly organized for waging conflict through combat. (Deutsch, 1969)

Pruitt and Kim (2004) present a model where escalation operates through changes in three areas: perceptions, patterns of behavior, and societal structures. Where fewer interpersonal ties exist to counter negative stereotypes about the out-group and in-group and institutional incentives foster antagonism, people employ more severe actions or rhetoric against the "other" (Pruitt and Kim, 2004). As people witness severe actions or rhetoric, they develop a basis for mistrust, resulting in "confident negative expectations regarding another's conduct" (Lewicki et al., 1998, 439). These persistent confirmed negative expectations alter the nature of groups and the self-protective ways they engage, reinforcing competitive, defensive, apathetic, and combative norms for interaction. Simplification abounds as complex issues are collapsed into simplistic truths and signals of group membership, with the resulting perception being that "instead of dealing with a particular threat from Other, Party must now deal with the general issue of how to resist an immoral enemy" (Pruitt and Kim, 2004*).* This perception of the other side as immoral and threatening paves the way for the remaining transformations that complete the escalation to violence.

A related body of conflict research is framed around "polarization," a broad concept which has been defined in many different ways (Bramson *et al.* 2017). Recent work in psychology and political science (Iyengar et. al, 2019) distinguishes between issue-based polarization, defined as the distance between parties on questions of policy, and relationship-based or affective polarization, meaning the increasing dislike, distrust, and animosity towards those from other parties or groups.

Just as some conflict can be constructive, issue-based polarization is not necessarily problematic. In contrast, affective polarization can increase the risk of escalation to violence by taking a conflict that is more specific and localized toward something more general, identity-based and antagonistic. Issue-based polarization becomes affective when we can't change what we think or say without losing core relationships or identities. Research on belonging and social boundaries points to an understanding that we are "driven not only by what we think, but also powerfully by who we think we are" (Mason, 2018). More broadly, conflict theorists consider increased polarization a warning sign for armed conflict (Esteban and Schneider, 2008; Laurenson, 2019) and the deterioration of democracy (McCoy and Somer, 2019).

However escalation is described, the end point of such a destructive spiral is either a tipping point where all parties are hurting so much that structural change becomes possible, or settling into a state of "intractable conflict" (Burgess and Burgess, 2023) where structures become rigid and de-escalation becomes very difficult.

Constructive and destructive conflict

Conflict is not inherently bad; it is part of how societies change for the better, and is sometimes necessary to achieve justice; it is an essential part of democratic debate, and necessary to hold power to account. In the words of Coser (1956): "Conflict prevents the ossification of the social system by exerting pressure for innovation and creativity."

Conflict scholars and political theorists have developed a variety of ways of talking about the dual nature of conflict. Deutsch (1969) talks of "constructive" and "destructive" conflict, noting that, for example, two parties can disagree about methods while agreeing on goals. Moufe (2013) distinguishes "agonistic" vs. "antagonistic" approaches to politics. McCoy and Somer (2019) are concerned with the effects of "pernicious" polarization on democracies. Political scientists talk about issue-based and affective polarization (Iyengar et. al, 2019). Sociologists investigate whether a social movement brings people together or tears them apart (Coley, Raynes and Das, 2020). Violence is a particularly extreme and destructive type of conflict, with lasting consequences; nonetheless philosophers have argued for millennia over the possibility of a "just war." Conversely, it is widely recognized that the mere absence of violence may hide deeper problems, leading to the concept of a "just peace" (Clements, 2004).

One fundamental difference between constructive or agonistic conflict and destructive or antagonistic conflict is how we feel about others when we take sides: when I hold an agonistic opinion, I disagree with you, but recognise your humanity and dignity; when I hold an antagonistic opinion my disagreement strips you of humanity or dignity. The theory of "agonistic democracy" recognizes that political factions often have fundamentally incompatible goals, and claims this conflict is not to be eliminated (for example, through partisan victory or authoritarian pacification) but transformed. Agonistic conflict is central to democracy; antagonistic conflict can destroy it:

> The aim of a pluralist democracy is to provide the institutions that will allow [conflicts] to take an agonistic form, in which opponents will treat each other not as enemies to be destroyed, but as adversaries who will fight for the victory of their position while recognising the right of their opponents to fight for theirs. An agonistic democracy requires the availability of a choice between real alternatives. (Mouffe 2000)

Many peacebuilding professionals subscribe to a related framework of “conflict transformation” that sees conflict, especially recurring cycles of conflict, as embedded in deeper structural problems, including systemic injustices. Conflict transformation seeks not to eliminate conflict but to change its nature (Lederach, 2003; Lederach, 2014; Clements, 2004). Mouffe similarly contends that the goal of democracy is to turn antagonistic conflict between “enemies” into agonistic conflict between “adversaries.” These ideas – and the corresponding practice of those professionals who must actually defuse violence – provide an important framework for intervening in conflict dynamics on social media.

While no definition of “good” versus “bad” conflict can account for all the richness of real conflict dynamics, in this paper, we will use the terms “destructive conflict” to refer to antagonistic conflict between affectively polarized opponents and “constructive conflict” to refer to agonistic conflict about issues.

Social media and conflict escalation dynamics

Conflict escalation is a long-term process, accompanied by negative changes in society long before the appearance of violence. Arguably, these changes are themselves harmful, but even the limited goal of preventing physical violence requires attention to conflict processes at far earlier stages. In this paper we are primarily concerned with how social media can manage and de-escalate conflict during ongoing operations, rather than only responding to crises where violence erupts. This dovetails with wider calls to develop the field of conflict prevention as a potentially much more effective and far less costly approach to managing conflict (United Nations Security Council, 2019).

An understanding of conflict escalation dynamics allows an analysis of the role of platforms in escalating destructive conflicts, and suggests ways they could be designed to de-escalate conflict. Escalation is a human process, but the architecture of social media platforms can amplify existing conflict dynamics, exacerbating fault lines and reinforcing destructive patterns of behavior (Puig Larrauri and Morrison, 2022).

Yet escalation doesn't automatically equal violence. If the structures of society contain safeguards (strong institutions, rule of law, legitimate and trusted conflict resolution systems, etc.) then there is less risk of wide-spread violence (Kriesberg, Louis and Dayton, 2012; Lederach, 1997). Platforms, as one of the major mediators of both public and private communication, have a role to play in conflict resilience. At the very least, they should not create additional risk by amplifying destructive conflict escalation cycles. At best, they should create the enabling conditions for constructive conflict to unfold.

Certain types of conflict actors are primarily financially motivated, as we will see below, and platforms should not allow such actors to inflame broader divisions. It is more difficult to judge politically-motivated conflict. Mass social movements universally claim to be fighting for justice, and exploiting pre-existing divisions is an effective political strategy (McCoy and Somer, 2019). How then should platforms react to polarizing strategies? One answer is to judge movements by the goals they espouse; but the means also matter, and anyway platforms are not equipped to make global judgments of who is in the right – nor should we grant them such power. Yet suppression of all conflict is authoritarian pacification, while universal support just allows conflicting parties to escalate unchecked. We argue that the correct goal of social media design is neither to eliminate conflict nor to judge the merits of specific parties, but to incentivize constructive over destructive conflict.

In the remainder of this article, we examine how the current design of social media often increases incentives towards destructive conflict and reduces incentives towards constructive conflict, and what can be done about it.

The relationship between social media and conflict

The most comprehensive reviews of the relationship between social media and constructs like “polarization” suggest a positive correlation (Kubin *et al.* 2021, Lorenz-Spreen *et al.*) The question of causation is more complex, and requires a deeper analysis of several plausible causal mechanisms and a variety of relevant evidence. Many of these questions center around the recommender systems that algorithmically select content for each user, because one of the core questions of conflict-sensitive platform design is who is exposed to what.

Filter bubbles are probably not driving polarization

The “filter bubble,” “echo chamber,” and “rabbit hole” metaphors encompass a variety of hypotheses about the possibility of narrow or one-sided exposure to information. These metaphors have been central to discussions of the relationship between social media and polarization for the last decade. If these hypotheses are true, then polarization could be reduced by increasing exposure to counter-ideological content (Stray 2022).

However, the accumulated evidence does not support the idea that filter bubbles are driving increases in polarization, at least for most users. Social media has been found to broaden the information diets of most users (Barbera, 2020). The divisions that exist on platforms generally pre-date social media (Boxell 2020). Further, increasing exposure diversity on social media (by asking people to follow a counter-ideological news source) may only have small effects on polarization (Stray 2022) or in some cases, can even make polarization worse (Bail *et al.* 2018).

Meta-analyses of the positive effects of inter-group contact suggest that it is not mere exposure to the outgroup that produces change, but rather the quality of that exposure (Pettigrew & Tropp 2006) including factors such as a cooperative environment, common goals, equal status, and norms endorsing contact. Clearly, many online interactions with alternative viewpoints do not meet these criteria, suggesting possible reasons why the mere exposure to counter-attitudinal information online does not have the desired effect.

There is a related family of theories about “rabbit holes,” the idea that recommender systems are making people more extreme as a result of a feedback loop between user beliefs and recommender outputs (Thorburn, Stray & Bengani 2023). An effect of this nature appears in certain stylized simulations of recommender operation (Mansoury et al. 2020, Carroll et al. 2021). Some studies of YouTube show this effect using bots that click randomly (Brown *et al.* 2022). However, users do not click randomly so this approach greatly overestimates rabbit hole effects (Ribeiro *et al.* 2023), and users don’t watch extreme videos on YouTube more than they consume them across the broader web (Hosseinmardi *et al.* 2021), which suggests a limited causal role for YouTube’s recommender.

One notable limitation of these studies is that they generally focus on average effects, whereas studies of radicalization often focus on individuals who commit extreme acts (e.g. Koehler, 2014; Roose, 2019). Foe these more extreme individuals, there are typically both online and offline processes at play (Gill *et al.* 2017, Baugut and Neumann 2020) suggesting that processes may be longer-term and involve ecosystem-level effects. Generally, we believe that there are more widespread and reliable phenomena than “filter bubbles” and “rabbit holes” for conceptualizing the relationship between social media, polarization, and conflict.

Social media’s broader negative impact on conflict dynamics

While criticisms based on filter bubbles and rabbit holes may exaggerate short-term impact on the average person, there remain areas where social media does impact the broader population and therefore has a responsibility for conflict outcomes.

There have always been actors who deliberately escalate conflict by heightening the divisions between groups. These have been called “conflict entrepreneurs” (Friis 1999, Ripley 2021) or “political entrepreneurs” (McCoy *et al.* 2019). Escalating conflict tends to be more destructive when the motivations of actors are more about furthering their own goals, rather than achieving a societal benefit (Ripley, 2021). The efforts of such actors have been aided and amplified by the affordances of platforms. In addition, many actors who would otherwise refrain from divisive tactics have reported being pushed towards more antagonistic rhetoric, in order to receive increased distribution.

In this section, we lay out evidence for how social media platforms are impacting conflict escalation dynamics across the globe, leading to more destructive conflict. Note that we do not claim that social media is the primary driver of conflict, nor that the harms of social media outweigh the benefits which seem to include, for example, greater political knowledge and participation (Lorenz-Spreen et al., 2021). Further, many of the processes we identify have long existed in other forms of media, for example, the use of radio to escalate violence in Rwanda (Puig Larrauri and Morrison, 2022). Rather, we are saying that certain social media dynamics are negative externalities and significant drivers of destructive conflict, regardless of the relative contribution of other drivers of conflict or the good that social media may do in other domains.

The enabling of mass harassment and manipulation

Social media's open system enables individual untrusted actors to target individuals *en masse* without the offline constraints of privacy, negative feedback, and the need to protect their reputation. For business reasons, social media systems are often designed with public visibility as the default setting (Frenkel and Kang, 2021) and users on social media platforms may sign up for accounts without realizing that they are discoverable by strangers by default. This means that conflict actors can reach a wide array of targets, without the high economic costs or social consequences they would normally experience offline. Youth, the elderly, and particularly vulnerable individuals are usually afforded some protection from strangers by others in the community who mediate those interactions. Such protections (e.g. age appropriate design codes) are now being added retroactively to systems that were originally designed to be as frictionless and open as possible.

This role of social media in conflict escalation has been widely recognised by peacebuilding practitioners. In 2016, a UN panel of experts report on South Sudan concluded that "social media has been used by partisans on all sides, including some senior government officials, to exaggerate incidents, spread falsehoods and veiled threats or post outright messages of incitement" (United Nations Security Council, 2016). More generally, the UN's expert on human rights and freedom of expression stated that social media is fuelling hate speech in warzones creating an "extremely dangerous" situation for vulnerable civilians (United Nations, 2022). Below we discuss three broad strategies that conflict actors have used: sock puppets, misinformation, and targeted harassment

One strategy is to use a large number of centrally controlled accounts ("sock puppets") to create the appearance of a mass movement, and especially to manipulate recommendation algorithms into treating such content as genuinely popular. These accounts may be bots posing as humans or they may be individually operated by real people; either way they are used deceptively. There are so many examples – many uncovered by platform teams – that the phenomenon has a name in industry practice: "coordinated inauthentic behavior" (Cinelli *et al.*, 2022).

To take a few recent examples, a number of networks of inauthentic and hacked accounts on Twitter were found to be amplifying a narrative that Sudanese internet users opposed the government's decision to transfer al-Bashir to the International Criminal Court (Owen Jones, 2021). Later, a sock puppet network was found to be sharing content about the United Arab Emirates' support for and relationship with Sudan (Owen Jones, 2022). In both cases, these accounts were also involved in promoting inauthentic narratives in other Middle East countries. In Libya, coordinated networks have been used to bolster Khalifa Haftar's Libyan National Army (Grossman *et al.*, 2020) or to undermine UN-led attempts to forge peace (Stanford Internet Observatory, 2020). These networks have been shown to originate outside of Libya, notably in Egypt, the UAE, Saudi Arabia and Russia. In the Philippines, the Government has reportedly (per BuildUp's sources) used troll armies to push narratives critical of the Communist New People's Army and discredit a resumption of peace talks, mirroring other reports of the use of troll armies in the Philippines (Bengali & Harper, 2019).

Manipulation of information is another common approach (though it is important to note that falsehood is not required to mobilize people through divisive strategies, so eliminating misinformation would not eliminate destructive conflict). Users often have little indication of the original source of a piece of content and are therefore vulnerable to believing that content in their social feeds is trustworthy. Social proof is a powerful influence (Cialdini & Goldstein, 2004) and a small number of hyper engagers can push a narrative to make it seem popular to others, even when a wider silent majority disagrees. While the effects of Russian manipulation in the 2016 US election specifically may be exaggerated (Bail *et al.,* 2020; Eady *et al.,* 2023), the wider effects of intentional misinformation are likely broad. Small groups in India have been successful at pushing narratives blaming Muslims for various societal issues (Avaaz, 2019). A relatively small group of users was responsible for the rapid growth of the Stop the Steal movements in the US, based on the false premise of widespread electoral fraud (Tech Policy Press, 2023). In early 2021, Brazil's Federal Police reported that it had found evidence of "digital militias," an elaborate network of public officials — from the Federal Cabinet all the way to the municipality — creating inauthentic pages, posts, and comments on social media to produce fake news and attack democratic institutions (Global Voices, 2022). One of the authors has seen a rise in YouTube channels created specifically to share disinformation and/or pro-military content about Myanmar. Many of these channels are run by financially-motivated actors, who are creating disinformation in order to capitalize on YouTube's monetization options. These actors are primarily based in Cambodia and Vietnam, and some are also working to produce disinformation on Ukraine.

Targeted escalation can also take the form of harassment when platforms allow a small number of harassers to hyper-engage with great effect. Online harassment particularly impacts women, people of color, and minority groups, and often spills over into offline violence. During the recent Kenyan elections, Build Up found that hashtags were used in coordination by a small number of actors on Twitter to drown out the Kenya Kwanza conversation by targeting the party with #liefesto (Build Up, 2022). In Ethiopia, there have been reports that online trolls pose as members of different ethnic groups to incite tensions between them (Selegna, 2022). In 2014, a rumour spread on Facebook that a young Buddhist woman had been raped by two Muslim men in Mandalay, Myanmar. In response, a mob formed outside the teashop of the alleged attackers, sparking altercations that led to two deaths (Waheed, 2015).

Whether financially or politically motivated, these are just a few conflict-relevant examples of the widely-studied phenomenon of platform manipulation, much of which is polarizing or escalatory (King and Pan, 2022; Diresta *et al.*, 2020; Ong and Cabanes, 2018).Those who want to create destructive forms of conflict now have powerful new tools to aid in this effort, and the effectiveness of these tools means that some will adopt similar tactics, while others who might moderate the space, especially women (Krook & Sanin, 2020), may find it too toxic to engage (Anderson & Auxier, 2020). Because elections are generally zero-sum competitions, the effectiveness of inauthentic tactics means that opposing partisans will feel pressure to use them, leading to the proliferation of dark PR firms that offer disinformation for hire (Silverman, Lytvynenko, and Kung, 2020).

These are not new observations; and large platforms have made considerable investments in detecting coordinated manipulation and harassment campaigns, though not uniformly across the globe (e.g. Facebook, 2021). Smaller platforms may lack the resources, know how, or motivation. In either case, we argue that these problems should be understood as enabled by underlying design decisions. Reactive response will not be as effective in the long term as changes in the ways people can interact online. For example, WhatsApp has progressively reduced the number of groups that a message can be shared to at once, which has led to dramatic reductions in the spread of inflammatory rumors (Benton, 2022).

The incentive toward divisiveness for non-conflict actors

The design of platforms can not only benefit those seeking to intentionally divide others, but also influence those who would otherwise be more moderate. Evidence for the incentive toward divisiveness exists from three primary sources: the experiences of publishers, reports on experimental results from within platforms, and external studies of the relationship between engagement and indicators of divisiveness. In particular, most recommenders strongly favor items which the user is predicted to engage with in some way (Thorburn, Stray, Bengani 2022). Engagement is a useful signal of value to users, and essential in some form to any media business model. It is also an error-prone signal and attempting to maximize engagement can result in damaging side effects (Bengani, Stray and Thorburn 2022). In particular, if more engagement leads to greater distribution then content creators have an incentive to produce divisive content.

Many publishers, who do numerous experiments to understand what does or does not work to drive business relevant metrics, have reported this incentive toward divisiveness. Buzzfeed built their business on the systematic understanding of content performance leveraging frequent experimentation (Wang, 2017). Jonah Peretti, Buzzfeed's CEO, emailed Facebook in 2018 about the fact that the most divisive content they created was getting the most virality, creating an incentive to produce more of it. He specifically blamed an algorithm change that prioritized comments and reshares. Internal analyses in response to this email reportedly confirmed that "misinformation, toxicity, and violent content are inordinately prevalent among reshares." (Hagey & Horwitz, 2021). This same perverse incentive was noted by politicians in Europe (Morris, 2021), who called Facebook's ranking system a "hate algorithm" that deepened political polarization. Ben Sasse, a former US Senator who served on committees providing oversight of tech platforms, reported that many of the celebrities he had interviewed feel trapped by these incentives and that several who had tried to "break out of the vicious cycle of rage-inflammation" learned to "throw themselves back into the outrage loop" when "no one clicks" and "metrics plummet" (Sasse, 2018).

Convergent evidence for the incentives that publishers report can be found in experiments conducted by platforms that have been reported or leaked. Most public information indicates that predicted engagement is a major factor in content ranking for large platforms (Lada, Wang, Yan, 2021; Narayanan and Kapoor 2023; Oremus *et al.* 2021; Zhao *et al.* 2019). A recently reported Facebook change removed predicted comments and shares from the ranking formula for political content, and led small to reductions in platform usage (0.18% fewer visits) but also a greater than 50% decrease in “anger” emoji reactions as well as accompanying reductions in bullying, inaccurate information, and graphic content (Horwitz, 2023). Previous articles based on internal documents from Facebook have shown similar effects where, for example, changes away from engagement based ranking for health related content led to a 12% decrease in misinformation and a 7% decrease in negative interactions (Klepper & Seitz, 2021).

Leaked documents made available by Gizmodo (2023), which represent a small sample of the large number of experiments that platforms have done, show more convergent results where engagement based ranking relates to negative outcomes. In particular, they show that reducing the influence of predicted reshares in content ranking can reduce the spread of inflammatory content in at-risk countries (Anonymous, 2021), reducing the weight of downstream engagement leads to drops in misinformation prevalence (Anonymous, 2020a), reducing effect of anger reactions leads to reductions in misinformation and graphic content (Anonymous, 2020b), and that engagement incentives and measures of misinformation, graphic content, and bullying can tradeoff (Anonymous, 2019a). Taken together, the available evidence points to the existence of engagement based incentives within Facebook’s systems consistent with the described experiences of publishers, where more divisive content performs better. Twitter recently open-sourced its algorithm (Narayanan and Kapoor 2023) which revealed that Twitter similarly prioritizes content that it expects users to retweet and reply to, which means we might expect that similar conflict dynamics are playing out on that platform.

While external researchers are generally unable to do true experiments on platforms, analyses of public data and lab experiments have generated another line of evidence, showing the same relationship between engagement and divisive content. Much of this work has been on Twitter, where data has historically been more accessible. Studies using Twitter data have shown that moral-emotional language (Brady *et al.*, 2017; de Leon & Trilling, 2021) and outgroup derogation (Mercandante *et al.*, 2023; Rathje, Van Bavel, and van der Linden, 2021) are correlated with greater engagement.   An experiment conducted by external researchers comparing Twitter’s algorithmic feed to its chronological feed yielded similar results where algorithmically ranked political content was not only deemed more polarizing, but also lower quality (Milli et. al, 2023). Given these associations and the known platform optimization for engagement, it is unsurprising that publishers have reported an incentive toward divisive content.

Real world effects of mass harassment, manipulation and divisive content

One possible criticism of the above studies is that they measure reductions in the distribution of content thought to be divisive but do not measure conflict outcomes directly, for example through surveys assessing affective polarization or support for violence. Do divisive narratives really matter, and do they really lead to physical violence? Evidence that they do comes from both lab studies and the experience of peacebuilders.

The effects of various kinds of divisive content have been studied widely in psychology labs, where some of the most reliable ways to generate negative intergroup attitudes toward others are to manipulate fear (Riek et. al, 2006), use social influence (Turner, 1991; Mackie and Wright, 2023; Kim *et al.*, 2021), and create competition between groups (Diehl, 1990). Theoretical models backed up by experimental evidence have outlined the mechanisms by which "immersion in a realm of online hate speech" can progress to avoidance and discrimination, and eventually increase the likelihood of violence against outgroup members (Bilewicz and Soral, 2020). Critically, some studies find that people for whom digital media is a primary source of information about politics consider hate speech to be a social norm rather than delinquent behavior (Bilewicz and Soral, 2020), making contempt of outgroups socially acceptable, decreasing intergroup empathy, and paving the path to intergroup violence.

This is corroborated by the experiences of peacebuilders who have seen divisive material propagate widely, driving conflict escalation dynamics rooted in affective polarization (Hawke, 2022). Content about specific issues of contention is often drowned out by more general, simplified, and unspecified claims. This results in the silencing of moderate voices and the acceptance of influencers with high in-group validation, such that users from formerly neutral, adjacent, or cross-cutting positions accumulate into a limited number of camps with increasing in-group cohesion and polarized affiliations. As affiliation becomes more important, there is also a reduction in the quantity and quality of meaningful communication and everyday interaction that are normal to peaceful engagement.

Examples from the field illustrate this. In the run up to the 2022 elections in Kenya, the entry of former Nairobi Governor Mike Mbuvi Sonko into the Mombasa Gubernatorial race led to the emergence of online harmful content dividing Kenyans of Arab descent and non-Arab communities along the Kenyan Coast (Build Up and Search for Common Ground, 2022). A retweet network graph from this period shows three clear poles representing the three conflicting political parties. These relatively homogeneous sub-networks represent tight patterns of in-group content sharing, including many negative comments and hate speech about out-groups.

In Lebanon, a social media analysis confirmed the spread of Facebook posts and tweets attributing generalized blame for the country's shortcomings to Syrian refugees (Build Up, 2019). The posts and tweets occurred in tandem with increasing tension between refugee and host communities, as reported by multiple UN agencies. Interviews with civil society actors confirmed that the spread of such content was impacting attitudes among Lebanese towards Syrian refugees. The increased presence of hate speech impacted anti-discriminatory norms, normalizing the harassment and blame of Syrian refugees.

A forthcoming report by the Sudanese Development Initiative (SUDIA) found that conversations on Facebook and Twitter are an important factor in impeding a resolution of the political stalemate. Examining conversations around four key conflict topics, the report finds that politicians respond to opinions shared on social media in ways that suggest they assign as much importance to them as to offline realities.

This incentive toward divisive content exists outside of any individual and cannot be eliminated simply by removing oneself from social media. Thus, experiments that seek to isolate the effects of social media by testing what happens to people who stay off social media (e.g. Allcott *et al.,* 2020; Asimovic *et al.*, 2021) are unlikely to be able to measure the full effect on the conflict ecosystem. The incentive toward conflict will continue to operate on the publishers and politicians in a person's community, regardless of their individual usage of social media. The same incentives also apply to a person's friends and family, who will amplify messages from publishers and politicians, on and offline. This holds true even for contexts where a large proportion of the population is not directly connected to platforms. In South Sudan, peacebuilders found that hate speech spread on Facebook would reach people fighting on the frontlines who did not have access to the internet via a network of peers (Clifford, 2017).

Destructive conflict escalation enabled by social media affects society as a whole. To understand the broader effects, we need to move away from a paradigm of individual harms and towards collective harm – as that is what matters to peace.

Moderation is not enough to prevent conflict escalation

The fundamental weakness of moderation as a conflict management approach is that it addresses only the most obvious forms of hate speech, coordinated harassment, misinformation and incitement to violence, without considering the processes that escalate conflict to that point or the context that may make subtler forms of speech more likely to lead to violence (Dangerous speech Project, 2023). Emphasizing cultural practice, Udupa and Pohjonen (2019) urge us to move "beyond the binary and normative divisions of acceptable and unacceptable speech [and] pay attention to the everyday online practices that underlie contemporary digital cultures."

Furthermore, the attempt to use content moderation as a primary tool creates new negative effects, in the form of unfair over-enforcement and under-enforcement, backlash against perceived bias, and the censorship of important views (Douek, 2021). Notably, content moderation practice frequently rebounds on exactly those it is supposed to protect, including women and minorities (e.g. Dwoskin et. al, 2021). These effects work against any strategy that might de-escalate and transform conflict on platform.

Objective policies cannot capture dangerous speech

Trying to separate speech into "good" and "bad" faces a number of problems as a conflict management strategy. Dangerous speech – meaning speech that leads to violence – is often as much a product of the context and history in which it is said (Dangerous Speech Project, 2023), and evaluating such context is impossible within a scaled content moderation framework (Douek, 2021; Iyer, 2022). Technology companies themselves have noted that a great deal of harmful content approaches the border of "bad speech" without actually violating platform rules, and such "borderline" content receives more engagement even when users don't endorse it

(Zuckerberg, 2021). This can be mitigated to some extent by downranking borderline content, which many platforms do (Gillespie 2022), but this still requires complex judgments of which content is deserving of this treatment.

More fundamentally, it is not difficult to escalate conflict without violating platform policies on hate speech or incitement to violence, especially against groups that have experienced historic discrimination. Human rights scholars have documented several other types of speech that precede violence (Dangerous Speech Project, 2023) including expressions of fear and rhetoric around protecting children. Recent work has shown how fear based speech is often more prevalent than hate speech (Saha et. al, 2023; Saha et. al, 2021), and examples describing how online content leads to offline violence often describe fear speech (Taub & Fisher, 2018; Hegyi, 2020).

These kinds of speech cannot be captured by moderation policies because they are not inherently bad. Everyday activities such as the reporting of crime news can be linked to polarized attitudes (Peffley et. al 1996), but they are also important avenues to keeping oneself safe. Fear, collective emotion, and intergroup competition exist for adaptive social reasons and platforms justifiably point out that they reflect these basic human processes, which existed long before social media.

However, discussions of collective fear and competition were historically rare. The phrase “never cry wolf” illustrates the social cost of sparking fear, and the norms against using such techniques merely to attract attention. This has changed with the advent of new communications technology. For example, U.S. news headlines have come to express significantly more anger, fear, disgust and sadness in the last two decades (Rozado et al. 2022). Platforms are not responsible for the existence of fear-driven narratives that pit groups against each other, but rather for the incentivization and amplification of such content, and the resulting escalation dynamics.

Reliance on moderation leads to bias, censorship, and reactance

Aside from the difficulty in deciding which speech is “bad,” removing such speech is immediately troubling from a freedom of expression perspective, especially because this classification will always be incomplete and error-prone (Douek 2021). Since errors can never be made equal across languages, moderation across parties who speak different languages will always be biased toward one side or the other, and especially towards English and other colonial languages (e.g. BSR 2022). Even when enforcement thresholds are applied identically between groups, if those two groups have different base rates of violation then one group will be sanctioned more often (Mosleh *et al.* 2022) and there will also necessarily be a larger percentage of unwarranted removals (false positives) for people in that group (Chouldechova 2017). Removal may even inflame conflict by legitimating grievances, as an analysis of European right-wing extremism suggests (Ravndal, 2018).

Moreover, conflict scholars have already noted that the strategy of simply removing “bad speech” is likely to fail (Puig Larrauri and Morrison, 2022) because it does not engage with the underlying drivers of escalation. Professional conflict transformation practices do not operate by attempting to prevent people from speaking, even though it is understood that certain types of speech can escalate conflict. Peacebuilders and mediators take a “multi-partial” approach which aims to view the conflict from multiple perspectives, understand the interests of the different parties, and respect the dignity and humanity of everyone involved (Zhang, Bollen, Euwema 2020). Peacebuilding dialogues have to let everyone experience what it's like to be listened to, as this is key to eventually transforming the conflict in a more constructive direction. Removing actors or shutting down discourse can never be a systemic solution – conflict escalation will move elsewhere, to another platform, or take a different form.

Platforms could be designed to foster peace

The identification of the dynamics that push actors toward divisiveness and facilitate harassment and manipulation also points the way toward solutions. Some form of social media is likely to exist from now on, so it behooves us to improve upon these dynamics such that when conflict plays out online, it is not primarily a way to attract financially beneficial attention, retain power, or fuel violence. Having no conflict is unrealistic and unhealthy. Rather, the conflict which occurs should be productive, contained, and agonistic conflict which does not dehumanize the other. We organize conflict-sensitive platform design strategies into two broad categories: reducing destructive conflict and increasing constructive conflict.

Reducing the facilitation of destructive conflict

The links between platform operation and conflict escalation suggest a range of strategies beyond removing content. We discuss three: reducing engagement incentives to divisiveness, collecting additional feedback to discriminate between positive and negative engagement, and changing defaults to make it harder for conflict entrepreneurs to reach large numbers of people.

Strategy 1: reduce engagement incentives to divisiveness
The strategy with the most empirical support at this time is to reduce the weight of engagement signals in content selection, for those contexts where engagement has a tendency to incentivize divisive content. Some engagement interactions have no explicit user value judgment – users can comment, reply, retweet, spend time on or reshare content that they find objectionable or intriguing, but that they do not endorse. Using such ambiguous signals to control the distribution of material on sensitive topics is inherently risky, because it creates incentives toward conflict. Some platforms have already taken important steps to reduce such incentives. Notably, Facebook removed predicted comments and shares from the ranking formula for political content, resulting in greater than 50% decrease in anger reactions as well as accompanying reductions in bullying, inaccurate information, and graphic content (Horwitz *et. al*, 2023). We are aware of one other large platform which has taken similar steps.

Some categories of ranking signals might turn out to be too difficult to use in a conflict-sensitive manner. For example, using “time spent” as a ranking signal prioritizes content that is more attention-grabbing, so it may not be possible to use this signal in a way that does not also incentivize the production of divisive content. Facebook de-emphasized time spent as part of its meaningful social interactions change (Oremus, 2017), but it is unclear to what degree time spent still influences content ranking. Recent source code releases suggest it is not used at Twitter (Narayanan and Kapoor 2023), but we know time spent is a major signal for TikTok (Smith 2021), and is predicted by the YouTube recommender as well (Zhao *et al.* 2019). Algorithmic transparency efforts could attempt to definitively determine the influence of time spent and other ambiguous signals across platforms.

Other interactions and incentives may be more subtle or context dependent, and can be detected by monitoring the spread of types of material that can be reliably identified as divisive, or more destructive than constructive. Platforms could audit their algorithms to understand which design choices are leading to the incentive toward division that publishers have reported. Ideally, these audits would be public, and allow for visibility into the experimental results that platforms use to understand the impact of design choices.

Strategy 2: collect additional feedback to discriminate between positive and negative engagement

In addition to standard signals such as comments, shares, and time spent, other kinds of feedback signals might help users differentiate between content that is genuinely valuable, content they agree with, and content that they react to without necessarily endorsing.

A lab experiment with “like,” “recommend,” and “respect” buttons found that people were more likely to “respect” than “like” content they disagreed with (Stroud, Muddiman and Scacco 2017). Similar designs (e.g. an “informative” button) could help algorithms find and surface less divisive and more informative content. Conversely, platforms also ought to give users a prominent way to signal that content is of negative value, such as thumbs down, hide, or “see less” buttons, as such negative signals are important for moderating offline interactions and could similarly be useful online (Anonymous, 2019b).

In general, the problem of determining whether engagement means an item is genuinely valuable or merely attention-getting requires the collection of some sort of additional feedback, and there are many ways to do this including providing new user controls and directly asking a subset of users with surveys. Better conflict is one of many values we might want social media to support, and the methods to measure and operationalize these values are developing rapidly (Stray *et al.* 2022).

Strategy 3: change defaults to make it harder for conflict entrepreneurs to reach large numbers of people

The third anti-escalation strategy we advocate for is a shift away from global distribution by default. Rather than defaulting to a design where any user can contact any other user, platforms could better attempt to ascertain the privacy desires of their users and enable those choices; what is good for business is not necessarily a good default from a conflict perspective. Such functionality has already proven to be a useful tool in some countries (Saini, 2020), and these tools should be made more widely accessible.

Similarly, rather than allowing a new, untrusted user the power to impact a large group of strangers, platforms should mirror real-life processes whereby individuals have to gain some level of trust to be able to reach broad groups of others. For example, Facebook has successfully used reputation signals to limit virality, with benefits in terms of reducing misinformation (Rodriguez, 2019). It would be better for individual users, who would be less subject to harassment from swarms of untrusted and potentially inauthentic users, and for society as a whole, if individuals who get broad distribution first need to earn some level of trust in the broader community.

Increasing incentives towards constructive conflict

Beyond reducing the facilitation of destructive conflict, platforms could be designed with constructive conflict in mind. The overall goal of this work would be to direct conflict in more constructive directions (Deutsch, 1973) rather than to suppress it entirely, in line with conflict transformation practices (Lederach, 2003; Lederach, 2014). From a peacebuilding perspective, this is about promoting positive, constructive, cross-cutting encounters (Pettigrew & Tropp 2006), cross-cutting group affiliations (Gaertner et. al, 1999), more complex and diverse narratives (IFIT 2021), and more complex, nuanced voices that model empathy and curiosity as norms. A number of studies have shown how important norms formed by example are in human behavior generally (Gelfand & Harrington, 2015) and in the online world specifically (Berry & Taylor 2017; Bilewicz and Soral, 2020). We discuss three concrete strategies that could connect these principles to social media systems: algorithmically promoting bridging content, exposing people to alternative content, and conducting (and possibly automating) moderating encounters.

Strategy 4: algorithmically promote bridging content

Bridging-based ranking prioritizes content that meets approval (or generates positive engagement) across diverse groups of people. This approach attempts to counteract the amplification of divisive material by favoring items which have cross-partisan appeal (Ovadya and Thorburn 2023). A simple example is Facebook's use of a crowdsourced survey to rate the credibility of news domains, rating as trustworthy only those with a supermajority of support (Owen, 2018). Twitter's Community Notes system (formerly Birdwatch), which asks users for crowdsourced notes on misleading tweets, is a much more sophisticated approach. Raters rank multiple notes, and this user-note rating matrix is factored to separate out high ratings due to partisan agreement from high ratings due to overall note quality. Only those notes which are widely agreed to be high quality are displayed with the original tweet (Wojcik *et al.* 2022). Bridging is also the core idea of Polis, a successful deliberative democracy system that collects and clusters opinions on political issues, mapping the points of consensus (Small *et al.* 2021).

There are many potential ways to identify bridging content. Local peacebuilders in Build Up's network have suggested allowing users to flag accounts which promote positive interaction or peace messaging. Promoting content which models constructive conflict is only possible if such content already exists on the platform. However, such promotion could change the incentives for the production of this type of bridging content, just as current engagement optimization incentivizes divisive content.

Strategy 5: expose people to constructive content
Beyond highlighting user existing posts, it is also possible to foster constructive conflict by showing carefully designed messages. The Strengthening Democracy Challenge (Voelkel *et al.* 2023) systematically tested many different interventions (each of which had to be done online, alone, and in less than 8 minutes) and found that 23 out of 25 improved intergroup attitudes, including reducing partisan animosity and reducing support for partisan violence. The interventions that most effectively reduced partisan animosity did so by either highlighting sympathetic and relatable individuals with different political beliefs, or presenting group identities that were common across partisan lines. The interventions that most effectively reduced support for partisan violence did so by correcting misperceptions of outpartisans' views or providing pro-democratic cues from someone in the political elite. Understanding how design decisions may incentivize or disincentivize such content could help platforms make more conflict aware design choices.

Strategy 6: support moderating encounters
When peacebuilders work on platforms they act as guides, coaches and bridge builders. They connect social media users to conversations that otherwise wouldn't happen, expose them to other voices and resources, and attempt to shift discourse toward shared values of civility and respect. For example, The Commons project sought out Americans who were expressing polarizing views, and engaged them in a text conversation with the aim of providing a humanizing experience of communication without changing their opinion (Build Up, 2019a). This approach was adapted and replicated in Kenya by a coalition of six universities, with similarly positive results (Ogenga, 2022). In Sri Lanka, the Cyber Guardians project of Search for Common Ground worked with social media influencers to change youth attitudes towards hate speech (Katheravelu, 2020). This sort of human facilitation work cannot yet be automated, but platforms could support existing peacebuilding efforts by promoting their programs in contexts where divisive conversations are likely to escalate.

Platforms might also consider providing API access to support more ambitious conflict transformation approaches. For example, it is possible to use large language models to help people rephrase their statements more constructively in a politically charged conversation (Argyle *et al.* 2023). Just as we have automated spelling checks in most products today, one could imagine these sorts of automated conflict assistants integrated into social media platforms.

No single design change is going to address conflict escalation in all circumstances. Conflict transformation is complex, and requires a shift in daily practices that eventually builds to a shift in societal norms. The design changes we suggest in this section could together help change the norms prevalent on platforms, away from divisiveness, hate and fear, and towards plurality and empathy.

The challenge of metrics

As the above discussion suggests, there are many design changes that might alter the trajectory of conflict on social media. Unfortunately, theory alone cannot tell us which will work best. We must test different approaches and evaluate the results against some measure of constructive conflict.

This is illustrated by the process used to develop Twitter's Community Notes, which tested eight different note ranking algorithms against two survey measures: agreement with misleading tweets, and trust in the appended notes (Wojcik *et al.* 2022). While a bridging-based ranking algorithm will involve the calculation of some sort of bridging signal – perhaps the difference in engagement across the sides in a conflict, or a matrix factorization approach like Community Notes – these types of signals cannot directly tell us what we really want to know: has a design change helped move the conflict from destructive to constructive?

So far, the conflict-relevant changes that have been implemented at platforms have mostly been evaluated using metrics designed for content moderation, such as the number of posts containing hate speech, incitement to violence, or misinformation, and the number of angry reactions generated, the number of accounts suspended for rule violations, and other similar indicators. These all have relevance to conflict, but were not designed to measure conflict intensity, nor discriminate between constructive and destructive conflict. Incitement to violence does not capture pre-violent escalation. Hate speech is not necessarily escalatory, and much violence is not driven by hate but fear (Leader Maynard and Benesch, 2016; Taub and Fisher, 2018; Hegyi, 2020). Misinformation is often divisive, but it is only one aspect of conflict.

Many other measures might provide better information about the state of a conflict. The Strengthening Democracy Challenge (Voelkel *et al.* 2023) tested each intervention against eight indicators: partisan animosity, support for undemocratic practices, support for partisan violence, support for undemocratic candidates, opposition to bipartisan cooperation, social distrust, and social distance, and biased evaluation of politicized facts. One could also add measures for affective polarization, dehumanization, and others.

All of these are survey measures, which can provide considerably more information than on-platform behavior alone. For example, Facebook asked users whether they perceived particular items to be “bad for the world” (Pawha, 2021; Anonymous, 2020c) which tended to be a signal of posts which were highly engaging yet more likely to contain hate speech, incitement, or graphic violence. Highly reshared content was more likely to be judged by users to be “bad for the world” (Anonymous, 2020c). This is an admittedly imperfect but potentially useful signal as to whether on-platform conflict is getting better or worse. Still, survey measures can be limited by user subjectivity and sample size, and so ideal measurement would combine methodologies across survey, content, and engagement modalities to mitigate the error of any one method (see Stray *et al.*, 2022 for a discussion).

Ideal metrics would be public facing and previously agreed upon by external stakeholders (Stray 2020). This is both a democratic and a pragmatic concern, as platforms may perceive no incentive to invest in conflict mitigation if they expect to be criticized regardless of anything they do. Such metrics could be used by researchers, regulators, advertisers, and the general public to hold platforms accountable for their design decisions in a way that is not currently possible. No metric is perfect, but an imperfect metric can be helpful, as long as it is not strongly optimized for (Manheim and Garrabrant 2018, Zhuang and Hadfield-Menell 2020).

In the final analysis, it is global society, not platforms, who must decide on how we evaluate conflict, including how we measure whether it is constructive or destructive.

Barriers to implementation

If platforms have made earnest efforts to improve their relationship to conflict, why do the experiences of those within conflict settings still suggest that the net effect is negative? One answer is that there are structural barriers that exist within large platforms and the business incentives they experience that may make progress difficult.

When it is easy to measure business outcomes and hard to measure societal impact, the basic desire to reduce cognitive dissonance will lead even the most well-meaning business to assume their business metrics are not at odds with societal needs. The complexity of the problem also means that there are few widely agreed-upon metrics that disambiguate constructive from destructive conflict. It will not be possible to create good metrics without the data, experimental capability and deep operational knowledge that platforms possess, yet the process of creating and legitimating a metric must also involve external stakeholders (Stray 2020).

Beyond creating public metrics, society should help platforms by taking some of the complex decision making out of their hands. Just as building designers have clear guidelines as to what safety standards are expected of them from society, so too could society provide clear guidance to companies as to what design patterns they need to follow. The design strategies above are informed by previous work. New research could help uncover other design patterns that could eventually be incorporated into conflict-sensitive design principles for online spaces. Some part of that research will inevitably (and sometimes necessarily) be done within companies, and it is hoped that companies, academics, policymakers, and engaged citizens could eventually work together to incorporate that evidence into our overall body of knowledge. Currently, collaborations with external researchers are very difficult to arrange, but we hope that forthcoming regulation will improve that, such as the researcher data access provisions of the EU Digital Service Act.

Conclusion

There is now good evidence, from multiple methods and perspectives, that social media platforms have had negative effects on societal conflict by pushing moderate actors toward divisiveness and enabling the actions of conflict entrepreneurs. These problems cannot be solved by content moderation, but must be addressed through design changes that help prevent the escalation of destructive conflict. From all of this evidence and experience, we have identified six broad strategies platforms might use to discourage destructive conflict before it escalates to violence.

1. **Reduce engagement incentives to divisiveness.** Reduce the weight of engagement signals in content selection, for those contexts where engagement has a tendency to incentivize the production of destructive conflict.

2. **Collect additional feedback to discriminate between positive and negative engagement.** New kinds of reactions (e.g. an “informative” button), controls, and user surveys might help distinguish between attention and value.

3. **Change defaults to make it harder for conflict entrepreneurs to reach large numbers of people.** Shift away from global distribution by default, and rely more on community and reputation.

4. **Algorithmically promote bridging content.** It’s not just engagement that matters, but the diversity of the people who are engaging.

5. **Expose people to constructive content.** Professional peacebuilders produce a wide variety of media designed to transform destructive conflict, and experimental evidence confirms that it shifts attitudes.

6. **Support moderating encounters.** Find ways to help people have positive online encounters, including API-level integration with peacebuilding programs that aim to connect people.

To their credit, platforms have taken some of these steps toward improving their impact on conflict that we can learn from and build upon. Ample evidence exists for a design playbook for platforms to improve their relationship to conflict, and society has an active role to play in partnering with platforms on the maintenance of that playbook and measurement of results.

References

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629-676.

Anderson, M., & Auxier, B. (2020, September 2). *55% of U.S. social media users say they are 'worn out' by political posts and discussions*. Pew Research Center. Retrieved April 17, 2023, from https://www.pewresearch.org/fact-tank/2020/08/19/55-of-u-s-social-media-users-say-they-are-worn-out-by-political-posts-and-discussions/

Anonymous (2021). *Big Levers Ranking Experiment.* Gizmodo Facebook Papers Directory. Retrieved April 17, 2023, from https://www.documentcloud.org/documents/21602350-tier0_rank_ro_undated

Anonymous (2020a). *[Launch] Replacing share downstream value for Civic and Health.* Gizmodo Facebook Papers Directory. Retrieved April 17, 2023, from https://www.documentcloud.org/documents/21600494-tier0_rank_ro_0620

Anonymous (2020b). *[Launch] Using p(anger) to reduce the impact angry reactions have on engagement ranking levers.* Gizmodo Facebook Papers Directory. Retrieved April 17, 2023, from https://www.documentcloud.org/documents/21600494-tier0_rank_ro_0620

Anonymous (2020c). *How much of News Feed is Good (or Bad) for the world?* Gizmodo Facebook Papers Directory. Retrieved April 17, 2023, from https://www.documentcloud.org/documents/21748444-tier2_fa_news_ir_1020

Anonymous (2019a). *Max Reshare Depth Experiment.* Gizmodo Facebook Papers Directory. Retrieved April 17, 2023, from https://www.documentcloud.org/documents/21602015-tier1_rank_pr_1119

Anonymous (2019b). *Providing Negative Feedback Should Be Easy (And Why This Would Be Game Changing For Integrity).* Gizmodo Facebook Papers Directory. Retrieved April 17, 2023, from https://www.documentcloud.org/documents/21602498-tier1_rank_pr_0919

Argyle, L. P., Busby, E., Gubler, J., Bail, C., Howe, T., Rytting, C., & Wingate, D. (2023). AI Chat Assistants can Improve Conversations about Divisive Topics (arXiv:2302.07268). arXiv. http://arxiv.org/abs/2302.07268

Asimovic, N., Nagler, J., Bonneau, R., & Tucker, J. A. (2021). Testing the effects of Facebook usage in an ethnically polarized setting. Proceedings of the National Academy of Sciences, 118(25). https://doi.org/10.1073/pnas.2022819118

Avaaz (2019). Megaphone for Hate. Retrieved April 17, 2023 from https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf

Bail, Christopher A. et al. 2018. “Exposure to Opposing Views on Social Media Can Increase Political Polarization.” PNAS 115(37): 9216–21.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency’s impact on the political attitudes and behaviors of American Twitter users in late 2017. Proceedings of the National Academy of Sciences, 117(1), 243–250. https://doi.org/10.1073/pnas.1906420116

Barberá, Pablo. 2020. “Social Media, Echo Chambers, and Political Polarization.” In Social Media and Democracy, eds. Nathaniel Persily and Joshua A. Tucker. Cambridge University Press, 34–55.

Baugut, P., & Neumann, K. (2020). Online propaganda use during Islamist radicalization. Information, Communication & Society, 23(11), 1570–1592. https://doi.org/10.1080/1369118X.2019.1594333

Bengali, S., & Harper, E. (2019, November 19). *Troll armies, a growth industry in the Philippines, may soon be coming to an election near you*. Los Angeles Times. Retrieved April 17, 2023, from https://www.latimes.com/politics/story/2019-11-19/troll-armies-routine-in-philippine-politics-coming-here-next

Bengani, Priyanjana, Jonathan Stray, and Luke Thorburn. (2022). “What’s Right and What’s Wrong with Optimizing for Engagement.” Understanding Recommenders. https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851.

Benton, Joshua. (2022). WhatsApp seems ready to restrict how easily messages spread in a bid to reduce misinformation. Nieman Journalism Lab. https://www.niemanlab.org/2022/04/whatsapp-seems-ready-to-restrict-how-easily-messages-spread-in-a-bid-to-reduce-misinformation/

Berry G. & Taylor S. (2017). Discussion Quality Diffuses in the Digital Public Square. Retrieved April 17, 2023 from https://dl.acm.org/doi/10.1145/3038912.3052666

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. Political Psychology, 41(S1), 3–33. https://doi.org/10.1111/pops.12670

Boxell, Levi, Gentzkow, Matthew, and Jesse M Shapiro. 2020. Cross-Country Trends in Affective Polarization. https://www.nber.org/papers/w26669.

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

Bramson, Aaron et al. 2017. “Understanding Polarization: Meanings, Measures, and Model Evaluation.” Philosophy of Science 84(1): 115–59.

Broockman, David, Joshua Kalla, and Sean Westwood. 2020. “Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not.” https://osf.io/9btsq/

Brown, Megan A. et al. 2022. “Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users.” https://papers.ssrn.com/abstract=4114905

BSR. 2022. Human Rights Due Diligence of Meta’s Impacts in Israel and Palestine. Retrieved April 14, 2023, from https://www.bsr.org/en/reports/meta-human-rights-israel-palestine

Build Up. 2019. Analyzing Refugee-Host Community Narratives On Social Media. UNDP Lebanon.
https://howtobuildup.org/wp-content/uploads/2020/06/UNDPBU_SocialMediaAnalysis_Leb_FINAL_310519.pdf.

Build Up, 2019a. The Commons: An intervention to depolarize political conversations on Twitter and Facebook in the USA.
https://howtobuildup.org/wp-content/uploads/2020/04/TheCommons-2019-Report_final.pdf

Build Up. (2022). Exploring Online Discourse in Kenya.
https://howtobuildup.org/wp-content/uploads/2022/11/Exploring-online-discourse-in-Kenya-Junefn.pdf

Build Up & Search for Common Ground. (2022). Uchaguzi Bila Balaa.
https://howtobuildup.org/wp-content/uploads/2023/02/Social-media-listening-analysis_final.pdf

Burgess, Heidi, and Guy M Burgess. 2003. “What Are Intractable Conflicts?” Beyond Intractability. https://www.beyondintractability.org/essay/meaning_intractability (accessed March 26, 2023).

Carroll, Micah, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. 2021. “Estimating and Penalizing Preference Shift in Recommender Systems.” RecSys ’21: Fifteenth ACM Conference on Recommender Systems. https://dl.acm.org/doi/10.1145/3460231.3478849.

Coser, Lewis. The Functions of Social Conflict. New York: The Free Press, 1956.

Chouldechova, Alexandra (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data 5(2). https://doi.org/10.1089/big.2016.0047

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. Annu. Rev. Psychol., 55, 591-621.

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on Twitter. Decision Support Systems, 160, 113819. https://doi.org/10.1016/j.dss.2022.113819

Clements, K. (2004). Towards Conflict Transformation and a Just Peace. In A. Austin, M. Fischer, & N. Ropers (Eds.), Transforming Ethnopolitical Conflict: The Berghof Handbook (pp. 441–461). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-663-05642-3_21

Clifford, Lisa. (2017, September 5). Words matter: Hate speech and South Sudan. The New Humanitarian. https://www.thenewhumanitarian.org/analysis/2017/09/05/words-matter-hate-speech-and-south-sudan

Coley, J. S., Raynes, D. K. T., & Das, D. (2020). Are social movements truly social? The prosocial and antisocial outcomes of social movements. Sociology Compass 14(8). https://doi.org/10.1111/soc4.12820

Dangerous speech: A practical guide (2023) *Dangerous Speech Project*. Available at: https://dangerousspeech.org/guide/ (Accessed: April 17, 2023).

de León, E., & Trilling, D. (2021). A sadness bias in political news sharing? The role of discrete emotions in the engagement and dissemination of political news on Facebook. Social media+ society, 7(4), 20563051211059710.

Deutsch, M. (1969). Conflicts: Productive and Destructive. Journal of Social Issues, 25(1), 7–42. https://doi.org/10.1111/j.1540-4560.1969.tb02576.x

Diehl, M. (1990). The minimal group paradigm: Theoretical explanations and empirical findings. European review of social psychology, 1(1), 263-292.

DiResta, R., Miller, C., Molter, V., Pomfret, J., & Tiffert, G. (2020). Telling China's Story: The Chinese Communist Party's Campaign to Shape Global Narratives. Stanford Internet Observatory.

Douek, E. (2021). Governing Online Speech: From “Posts-as-Trumps” to Proportionality and Probability. Columbia Law Review, 121(3). https://columbialawreview.org/content/governing-online-speech-from-posts-as-trumps-to-proportionality-and-probability/

Druckman, James N, and Matthew S Levendusky. 2019. “What Do We Measure When We Measure Affective Polarization?” Public Opinion Quarterly 83(1): 114–22.

Dwoskin, E., Tiku, N., & Timberg, C. (2021). Facebook's race-blind practices around hate speech came at the expense of Black users, new documents show. https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. Nature Communications, 14(1), 62. https://doi.org/10.1038/s41467-022-35576-9

Esteban, Joan, and Gerald Schneider. 2008. "Polarization and Conflict: Theoretical and Empirical Issues." Journal of Peace Research 45(2): 131–41.

Facebook, (2021 November). November 2021 Coordinated Inauthentic Behavior Report. https://about.fb.com/wp-content/uploads/2021/12/November-2021-CIB-Report.pdf

Finkel, Eli J. et al. 2020. "Political Sectarianism in America." Science, 370(6516): 533–36.

Frenkel, S., & Kang, C. (2021). An ugly truth: Inside Facebook's battle for domination. Hachette UK.

Friis, K. (2000). From Liminars to Others: Securitization Through Myths. Peace and Conflict Studies 7(2). https://doi.org/10.46743/1082-7307/2000.1008

Gaertner, S. L., Dovidio, J. F., Nier, J. A., Ward, C. M., & Banker, B. S. (1999). *Across cultural divides: the value of a superordinate identity*. Russell Sage Foundation.

Gelfand, M. J., & Harrington, J. R. (2015). The motivational force of descriptive norms: For whom and when are descriptive norms most predictive of behavior?. *Journal of Cross-Cultural Psychology*, *46*(10), 1273-1278

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. Social Media + Society, 8(3), 205630512211175. https://doi.org/10.1177/20563051221117552

Gizmodo. (2023, February 14). *Read the Facebook Papers for Yourself*. *Read the facebook papers for yourself*. Retrieved April 15, 2023, from https://gizmodo.com/facebook-papers-how-to-read-1848702919

Glazer, J. H., Keach Hagey and Emily. (2023). Facebook Wanted Out of Politics. It Was Messier Than Anyone Expected. WSJ. Retrieved January 31, 2023, from https://www.wsj.com/articles/facebook-politics-controls-zuckerberg-meta-11672929976

Global Voices (2022). Undertones: Brazil copes with ‘digital militias’ ahead of tense elections. Retrieved April 17 2023 from https://globalvoices.org/2022/10/27/undertones-brazil-copes-with-digital-militias-ahead-of-tense-elections/

Grossman, S. (2020, April 2). Blame it on Iran, Qatar, and Turkey: An analysis of a Twitter and Facebook operation linked to Egypt, the UAE, and Saudi Arabia (TAKEDOWN). FSI. https://fsi.stanford.edu/publication/twitter-facebook-egypt-uae-saudi

Hawke, Julie. (2022, April 30). Archetypes of Polarization on Social Media. Build Up blog. https://howtobuildup.medium.com/archetypes-of-polarization-on-social-media-d56d4374fb25

Hegyi, N. (2020) *The 'concerned citizen who happens to be armed' is showing up at protests*, *NPR*. NPR. Available at: https://www.npr.org/sections/live-updates-protests-for-racial-justice/2020/06/10/873751544/the-concerned-citizen-who-happens-to-be-armed-is-showing-up-at-protests (Accessed: April 17, 2023).

Horwitz, K. H. and J. (2021, September 15). Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. Wall Street Journal. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

Horwitz, J., Hagey, K., & Glazer, E. (2023, January 5). Facebook Wanted Out of Politics. It Was Messier Than Anyone Expected. Wall Street Journal. https://www.wsj.com/articles/facebook-politics-controls-zuckerberg-meta-11672929976

Hosseinmardi, Homa et al. 2021. “Examining the Consumption of Radical Content on YouTube.” Proceedings of the National Academy of Sciences 118(32)

IFIT (2021). The Role of Narrative in Managing Conflict and Supporting Peace. Institute for Integrated Transitions. https://ifit-transitions.org/publications/the-role-of-narrative-in-managing-conflict-and-supporting-peace/

Iyer, R. (2022). Content Moderation is a Dead End. [Substack newsletter]. The Psychology of Technology Institute Newsletter. https://psychoftech.substack.com/p/content-moderation-is-a-dead-end

Katheravelu, R. (2020). Cyber Guardians: Empowering youth to combat online hate speech in Sri Lanka. Inno Consulting Service. https://www.sfcg.org/wp-content/uploads/2020/05/SFCG-Sri_Lanka_Cyber_Guardians_Final_Evaluation_2020.pdf

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. Journal of Communication, 71(6), 922-946.

King, G., Jennifer Pan, and Margaret E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument." American Political Science Review, 111, 3, Pp. 484-501. Publisher's Version Copy at https://tinyurl.com/ybdarq39

Klepper, D., & Seltz, A. (2021, October 26). *Facebook froze as anti-vaccine comments swarmed users*. AP NEWS. Retrieved April 15, 2023, from https://apnews.com/article/the-facebook-papers-covid-vaccine-misinformation-c8bbc569be7cc2ca583dadb4236a0613

Koehler, D. (2014). The radical online: Individual radicalization processes and the role of the Internet. Journal for Deradicalization, (1), 116-134.

Kriesberg, Louis and Dayton, Bruce W. Constructive Conflicts: From Escalation to Resolution, Rowman & Littlefield 2012

Krook, M. L., & Sanín, J. R. (2020). The cost of doing politics? Analyzing violence and harassment against female politicians. *Perspectives on Politics*, *18*(3), 740-755.

Kubin, Emily, and Christian von Sikorski. 2021. "The Role of (Social) Media in Political Polarization: A Systematic Review." Annals of the International Communication Association: 1–19.

Lada, A., Wang, M., & Yan, T. (2021, January 26). *How does news feed predict what you want to see?* Tech at Meta. Retrieved April 15, 2023, from https://tech.facebook.com/engineering/2021/1/news-feed-ranking/

Laurenson, L, July 2019, "Polarisation and Peacebuilding Strategy on Digital Media Platforms", Toda Peace Institute Policy Brief No. 44, Accessed 19 February 2021, https://toda.org/assets/files/resources/policy-briefs/t-pb-44_laurenson-lydia_part-1_polarisation-and-peacebuilding-strategy.pdf

Leader Maynard, J., & Benesch, S. (2016). Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention. Genocide Studies and Prevention, 9(3), 70–95. https://doi.org/10.5038/1911-9933.9.3.1317

Lederach, John Paul. (2014). *The Little Book of Conflict Transformation.* Good Books.

Lederach, J.P., (1997), Building Peace: Sustainable Reconciliation in Divided Societies, United States Institute of Peace Press, Washington, DC

Lorenz-Spreen, Philipp, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. 2021. Digital Media and Democracy: A Systematic Review of Causal and Correlational Evidence Worldwide. SocArXiv. preprint. https://osf.io/p3z9v

Mackie, D. M., & Wright, C. L. (2003). Social influence in an intergroup context. Blackwell handbook of social psychology: Intergroup processes, 281-300.

Manheim, D., & Garrabrant, S. (2018). *Categorizing Variants of Goodhart's Law*. 1–10.

Mansoury, Masoud et al. 2020. "Feedback Loop and Bias Amplification in Recommender Systems." In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, New York, NY, USA: Association for Computing Machinery, 2145–48. https://doi.org/10.1145/3340531.3412152 (June 21, 2021).

Mason, Lilliana. 2018. "Ideologues Without Issues: The Polarizing Consequences Of Ideological Identities". Public Opinion Quarterly 82 (S1): 866-887. doi:10.1093/poq/nfy005.

McCoy, Jennifer, and Murat Somer. 2019. "Toward a Theory of Pernicious Polarization and How It Harms Democracies: Comparative Evidence and Possible Remedies." The Annals of the American Academy of Political and Social Science 681(1): 234–71.

Mercadante, E. J., Tracy, J. L., & Götz, F. M. (2023). Greed communication predicts the approval and reach of US senators' tweets. Proceedings of the National Academy of Sciences, 120(11), e2218680120.

Milli, S., Carroll, M., Pandy, S., Wang, Y, and Dragan, A. (2023). Twitter's Algorith: Amplifying Anger, Animosity, and Affective Polarization. Draft presented at Knight First Amendment Institute Algorithmic Amplification Symposium. https://knightcolumbia.org/events/optimizing-for-what-algorithmic-amplification-and-society

Mølmen, G. N., & Ravndal, J. A. (2021). Mechanisms of online radicalisation: how the internet affects the radicalisation of extreme-right lone actor terrorists. Behavioral Sciences of Terrorism and Political Aggression, 1-25.

More in Common, (2018). The Hidden Tribes of America.

Morris, L. (2021, October 27). In Poland's politics, a "social civil war" brewed as Facebook rewarded online anger. Washington Post. https://www.washingtonpost.com/world/2021/10/27/poland-facebook-algorithm/

Mosleh, Mohsen et al. 2022. Trade-Offs between Reducing Misinformation and Politically-Balanced Enforcement on Social Media. PsyArXiv. preprint. https://osf.io/ay9q5 (March 26, 2023).

Mouffe, Chantal. 2000. *The Democratic Paradox.* Verso.

Mouffe, Chantal. 2013. *Hegemony, Radical Democracy, and the Political.* Routledge.

Narayanan, A., & Kapoor, A. N. S. (2023, April 10). *Twitter showed us its algorithm. what does it tell us?* Knight First Amendment Institute. Retrieved April 15, 2023, from https://knightcolumbia.org/blog/twitter-showed-us-its-algorithm-what-does-it-tell-us

Northrup T. A. (1989). The dynamic of identity in personal and social conflict. In Kriesberg L., Northrup T. A., Thorson S. J. (Eds.), Intractable conflicts and their transformation (pp. 55–82). Syracuse University Press.

Ogenga, F. (2022). Maskani is Our New Normal- Exploring Digital Peacebuilding in Kenya, Working from Home. ConnexUs. https://cnxus.org/resource/maskani-is-our-new-normal-exploring-digital-peacebuilding-in-kenya-working-from-home-2020/

Ong, J. C., & Cabañes, J. V. A. (2018). Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines.

Oremus, W. (2017) *Facebook has a new philosophy. could it fix the Russia problem?*, *Slate Magazine*. Slate. Available at: https://slate.com/technology/2017/11/could-emphasizing-time-well-spent-fix-facebooks-russia-problem.html (Accessed: April 17, 2023).

Oremus, W., Alcantra, C. , Merrill, C, Galocha, A. WP Company. (2021, October 26). Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. The Washington Post. Retrieved April 15, 2023, from https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/

Ovadya, A., & Thorburn, L. (2023). Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance (arXiv:2301.09976). arXiv. https://doi.org/10.48550/arXiv.2301.09976

Owen, Laura Hazard. (2018). Crowdsourcing trusted news sources can work—But not the way Facebook says it'll do it. (n.d.). Nieman Lab. Retrieved April 15, 2023, from https://www.niemanlab.org/2018/02/crowdsourcing-trusted-news-sources-can-work-but-not-the-way-facebook-says-itll-do-it/

Owen Jones, M. (2021, August 13). *This thread is about a trend advocating for preventing Omar al-Bashir, wanted for crimes against humanity, from being sent to the ICC. It's a fantastic example of how artificially amplified and manipulated trends pretend to be grassroots sentiment #Disinformation #Sudan*. Twitter. https://twitter.com/marcowenjones/status/1426124574601658369

Owen Jones, M. (2022, January 14). *Thread 1/ For Sudan and Gulf watchers. Below is a brief analysis of a “Sudanese” sockpuppet network that includes at least 26 accounts. It seems to exist mostly to promote the UAE’s role in #Sudan, and occasionally have swipes at the Muslim Brotherhood. #disinformation*. Twitter. https://twitter.com/marcowenjones/status/1482056864552660997

Pahwa, N. (2021, November 15). Facebook asked users what content was "good" or "bad for the world." Some of the results were shocking. Slate Magazine. Retrieved April 1, 2023, from https://slate.com/technology/2021/11/facebook-good-bad-for-the-world-gftw-bftw.html

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. 90(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Peffley, M., Shields, T., & Williams, B. (1996). The intersection of race and crime in television news stories: An experimental study. *Political Communication*, *13*(3), 309-327.

Pruitt, D. G., & Kim, S. H. (2004). Social conflict: Escalation, stalemate and settlement (3rd ed.). New York: McGraw-Hill.

Puig Larrauri, H. and Morrison M. (2022) Understanding Digital Conflict Drivers, chapter in the book Fundamental Challenges to Global Peace and Security

Ribeiro, Manoel Horta, Veniamin Veselovsky, and Robert West. 2023. “The Amplification Paradox in Recommender Systems.” http://arxiv.org/abs/2302.11225 (February 24, 2023).

Riek, B. M., Mania, E. W., & Gaertner, S. L. (2006). Intergroup threat and outgroup attitudes: A meta-analytic review. Personality and social psychology review, 10(4), 336-353.

Ripley, A. (2021). High conflict: Why we get trapped and how we get out. Simon and Schuster.

Rodriguez, S. (2019) *Facebook is taking a page out of Google's playbook to stop fake news from going viral*, *CNBC*. CNBC. Available at: https://www.cnbc.com/2019/04/10/facebook-click-gap-google-like-approach-to-stop-fake-news-going-viral.html (Accessed: April 17, 2023).

Roose, K. (2019). The making of a YouTube radical. The New York Times, 8.

Rozado, D., Hughes, R., & Halberstadt, J. (2022). Longitudinal analysis of sentiment and emotion in news media headlines using automated labeling with Transformer language models. PLOS ONE, 17(10), e0276367. https://doi.org/10.1371/journal.pone.0276367

Ravndal, Jacob Aasland. 2018. “Explaining Right-Wing Terrorism and Violence in Western Europe: Grievances, Opportunities and Polarisation.” European Journal of Political Research 57(4): 845–66.

Saha, P., Garimella, K., Kalyan, N. K., Pandey, S. K., Meher, P. M., Mathew, B., & Mukherjee, A. (2023). On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, *120*(11), e2212270120.

Saha, P., Mathew, B., Garimella, K., & Mukherjee, A. (2021, April). “Short is the Road that Leads from Fear to Hate”: Fear Speech in Indian WhatsApp Groups. In *Proceedings of the Web Conference 2021* (pp. 1110-1121).

Saini, N. (2020, May 21) *Facebook now lets users 'lock their' profiles, here's how it works - times of India*, *The Times of India*. TOI. Available at: https://timesofindia.indiatimes.com/gadgets-news/facebook-now-lets-users-lock-their-profiles-heres-how-it-works/articleshow/75863324.cms (Accessed: April 17, 2023).

Sasse, B. (2018). Them: Why We Hate Each Other--and how to Heal. St. Martin's Press.

Selegna. (2022, January 26). Media Contents Censorships, Political influence, and Economic constraints. Selegna Media. https://selegnamedia.com/2022/01/26/media-contents-censorships-political-influence-and-economic-constraints/

Schirch, L. ed. (2021). Social Media Impacts on Conflict and Democracy: The Techtonic Shift. Routledge.

Senko, J. (2015) The Brainwashing of My Dad.

Silverman, C., Lytvynenko, J., & Kung, W. (2020, January 7). *Disinformation for hire: How a new breed of PR firms is selling lies online*. BuzzFeed News. Retrieved April 17, 2023, from https://www.buzzfeednews.com/article/craigsilverman/disinformation-for-hire-black-pr-firms

Small, C., Bjorkegren, M., Erkkilä, T., Shaw, L., & Megill, C. (2021). Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. RECERCA. Revista de Pensament i Anàlisi. https://doi.org/10.6035/recerca.5516

Smith, B. (2021, December 6). How TikTok Reads Your Mind. The New York Times. https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html

Stanford Internet Observatory. Stoking Conflict by Keystroke. (2020, December 15). FSI. https://cyber.fsi.stanford.edu/io/news/africa-takedown-december-2020

Stray, J. (2020). Aligning AI Optimization to Community Well-being. International Journal of Community Well-Being, 3, 443–463. https://doi.org/10.1007/s42413-020-00086-3

Stray, J. (2022). Designing Recommender Systems to Depolarize. First Monday, 27(5). https://firstmonday.org/ojs/index.php/fm/article/view/12604

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., Beattie, L., Ekstrand, M., Leibowicz, C., Sehat, C. M., Johansen, S., Kerlin, L., Vickrey, D., Singh, S., Vrijenhoek, S., Zhang, A., Andrus, M., Helberger, N., Proutskova, P., … Vasan, N. (2022). Building Human Values into Recommender Systems: An Interdisciplinary Synthesis (arXiv:2207.10192). arXiv. https://doi.org/10.48550/arXiv.2207.10192

Stroud, N. J., Muddiman, A., & Scacco, J. M. (2017). Like, recommend, or respect? Altering political behavior in news comment sections. New Media & Society, 19(11), 1727–1743. https://doi.org/10.1177/1461444816642420

Taub, A. and Fisher, M. (2018) *Where countries are tinderboxes and Facebook is a match*, *The New York Times*. The New York Times. Available at: https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html (Accessed: April 17, 2023).

Tech Policy Press. (2023, January 6). *Results of the january 6th Committee's Social Media Investigation*. Tech Policy Press. Retrieved April 17, 2023, from https://techpolicy.press/results-of-the-january-6th-committees-social-media-investigation/

Thorburn, L., Stray, J., & Bengani, P. (2022). How Platform Recommenders Work. Understanding Recommenders. https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a

Törnberg, Petter. 2022. “How Digital Media Drive Affective Polarization through Partisan Sorting.” Proceedings of the National Academy of Sciences 119(42): e2207159119.

Tufekci, Z. (2018). TWITTER AND TEAR GAS : the power and fragility of networked protest. Yale University Press.

Twitter. (2021, January 8). Permanent suspension of @realDonaldTrump. Blog.twitter.com; Twitter. https://blog.twitter.com/en_us/topics/company/2020/suspension

Udupa, S., & Pohjonen, M. (2019). Extreme Speech and Global Digital Cultures—Introduction. International Journal of Communication, 13. https://ijoc.org/index.php/ijoc/article/view/9102

United Nations. Digital technology, social media fuelling hate speech like never before, warns UN expert. (2022, October 20). UN News. https://news.un.org/en/audio/2022/10/1129712

United Nations Security Council, 2016, ‘Interim Report of the Panel of Experts on South Sudan Established Pursuant to Security Council Resolution 2206’, Accessed 16 February 2021, https://www.undocs.org/S/2016/963

United Nations Security Council (2019). More Unified, Early Action Key for Preventing Conflict, Reducing Human Suffering, Speakers Tells Security Council, Pointing to High Cost of Managing Crises, SC/13837. https://press.un.org/en/2019/sc13837.doc.htm

Voelkel, J. G., Stagnaro, M., Chu, J., Pink, S. L., Mernyk, J. S., Redekopp, C., Ghezae, I., Cashman, M., Adjodah, D., Allen, L., Allis, V., Baleria, G., Ballantyne, N., Van Bavel, J. J., Blunden, H., Braley, A., Bryan, C., Celniker, J., Cikara, M., … Willer, R. (2023). Megastudy identifying effective interventions to strengthen Americans' democratic attitudes [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/y79u5

Waheed, A. (2015, October 28). Rape used as a weapon in Myanmar to ignite fear. Humanitarian Crises . Al Jazeera. https://www.aljazeera.com/%20features/2015/10/28/rape-used-as-a-weapon-in-myanmar-to-ignite-fear

Wang, S. (2017) *BuzzFeed's strategy for getting content to do well on all platforms? adaptation and a lot of A/B testing*. Nieman Lab. Retrieved April 15, 2023, from https://www.niemanlab.org/2017/09/buzzfeeds-strategy-for-getting-content-to-do-well-on-all-platforms-adaptation-and-a-lot-of-ab-testing/

Warofka, A. (2018). An independent assessment of the human rights impact of Facebook in Myanmar. Facebook Newsroom, November, 5.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M. B. F., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation (arXiv:2210.15723). arXiv. https://doi.org/10.48550/arXiv.2210.15723

Wong, Q. (2023) Facebook temporarily blocked in Myanmar after military coup. CNET. Retrieved March 1, 2023, from https://www.cnet.com/news/politics/facebook-temporarily-blocked-in-myanmar-after-military-coup/

Zhang, Xiaolei, Katalien Bollen, and Martin Euwema. 2020. “Peacemaking at Work and at Home.” International Journal of Conflict Management 31(5): 801–20.

Zhao, Z., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., Kumthekar, A., Sathiamoorthy, M., Yi, X., & Chi, E. (2019). Recommending What Video to Watch Next: A Multitask Ranking System. RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems, 43–51. https://doi.org/10.1145/3298689.3346997

Zhuang, S., & Hadfield-Menell, D. (2020). Consequences of Misaligned AI. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. https://doi.org/10.5555/3495724.3497046

Zuckerberg, M. (2018). A Blueprint for Content Governance and Enforcement. Facebook, May 5. https://www.facebook.com/notes/751449002072082/