CHAPTER 9

Understanding Digital Conflict Drivers

Helena Puig Larrauri and Maude Morrison

This chapter explores digital drivers of conflict. We examine how technologies are affecting conflict dynamics and what peacebuilders can do to mitigate these effects. We argue that because digital technologies are fundamentally altering the human experience, they are by extension fundamentally altering conflicts. We propose a framework for understanding the impact of technology on conflict, and for categorising different types of peacebuilding interventions. Together, these interventions contribute to the emerging field of digital peacebuilding.

H. Puig Larrauri
Build Up, London, UK
e-mail: helena@howtobuildup.org

M. Morrison (✉)
Centre for Humanitarian Dialogue, Geneva, Switzerland

H. Mahmoudi et al. (eds.), *Fundamental Challenges to Global Peace and Security*,
https://doi.org/10.1007/978-3-030-79072-1_9

Part I: Introduction

Over the past 15 years, the peacebuilding field has begun to recognise the importance of digital technology—both in fuelling conflict and in supporting peacebuilding work. Initially, this led to the emergence of the term 'peacetech', referring to a growing body of peacebuilding practice that deployed technology as part of its strategic objectives. Across the globe, peacebuilders began to recognise that technology could enhance the impact of the work they had been doing for decades.

During the early phases of 'peacetech' work, many practitioners approached technology as neutral, as a tool that can be used for positive or negative effect depending on how we choose to use it. As Margot Wallstrom (2015, p. 36) noted in a 2015 issue of Building Peace dedicated to Peacetech, 'technology in itself is neutral and can be used for both good and evil'.

This idea was supported by numerous examples of both positive and negative uses of technology. As positive examples, people pointed to ways technology was enabling more or different voices to participate in discussions about peace, allowing new stories to surface and creating new opportunities for connection. At the same time, it was becoming clear that technology does not automatically lead to peace or positive social change. Negative and violent uses of technology were not hard to find: from video games promoting a culture of violence, to recruitment into armed groups through social media, the use of messaging apps to spread hate speech, and online surveillance by authorities. This dual use led to many viewing technology as a 'double-edged sword' (Youngs 2014). The narrative of peacebuilding and technology was one of a binary choice between threat and opportunity.

However, this vision of technology as a neutral tool ignored a key complexity—the fact that technology is fundamentally altering the human experience. Today, technology is a mediator of our experience of reality and to approach it as external to the conflict context is to miss the dynamics it fuels regardless of negative intent. In other words, technology is not just a tool that can be used to fuel the flames of violence or to dampen them, depending on the intention of the user. Instead, technology should be considered an integral part of the context in which conflicts occur, and peacebuilding responses should recognise and address these technological factors.

Before we expand our argument, two caveats. First, we are not claiming that technology is inherently *bad* because it affects our human experience. Indeed, it can still have positive and negative effects. Technology can divide communities or bring them together. It can provide incredible opportunities for innovation and equally incredible opportunities for destruction. The dual nature of its effects are not disputed. We dispute only the view that technology is a tool external to a context, and can be addressed as separate from the underlying causes of a given conflict.

Second, this chapter is not focused on cyber warfare and the use of digital tools as weapons of violence. The deliberate use of technology to inflict harm (for example through the use of drones or cyberattacks on infrastructure) is a distinct topic with a different set of challenges. The responses to these challenges are also different to those we propose here. Those responses are less likely to involve community peacebuilding and more focused on diplomatic efforts, such as the Digital Peace Now campaign which is working to outlaw state-sponsored cyberattackscyber attacks through an international agreement.[1] We will not address these issues in this chapter.

Instead, this chapter proposes a framework for peacebuilders to understand how technological factors are fundamentally altering (and driving) conflict in certain ways. We examine the ways in which technology creates the *enabling conditions* for conflict drivers that increase societal division, erode social cohesion, and amplify polarization—and can eventually lead to violent conflict. By better understanding this socio-technological context, peacebuilders can begin to think of digital technologies as a space for peacebuilding action: there is a need for conflict prevention and transformation in the digital space that addresses digital conflict drivers.

Part 2: Framework

The potential for positive uses of technology to address conflict has been well documented. Years ago, Puig Larrauri and Kahl (2013) published a framework outlining the functions that technology can play in peacebuilding—data processing, communication, engagement, and gaming.

[1] The Digital Peace Now Initiative is a global movement calling for an end to cyber warfare https://digitalpeacenow.org/about-us/.

Since then, that framework has been revised and updated as countless examples of peacebuilders using technology have emerged.[2]

In this chapter, we repurpose that framework to explore the role that technology can play in driving conflict. We posit that there are three core ways in which technology interacts with a context, creating the enabling conditions for conflict. They can be categorised as 'affordances', or what technology enables one to do. Across each of these core affordances, technology has the potential to alter our experience, and to shape conflict.

The three core affordances are:

- *Strategic communications*: The use of digital technologies to ***create and spread divisive content***, such as hate speech, misinformation and disinformation.
- *Data management*: The use of digital technologies to ***target and accelerate the spread of divisive content.*** This is currently most evident through algorithmic profiling, deliberate targeting and surveillance.
- *Networking*: The use of digital technologies to ***generate network effects that continue to drive communities apart.*** This is currently most evident through affective polarization, divisive identity formation and recruitment into violence.

The three categories of affordances are interconnected. The strategic communications affordance refers to the *content* that can drive conflict. The data management affordance refers to a set of *tools* that enable conflict actors to more effectively use that content to sew division. The networking affordance creates a set of *enabling conditions* that make the combination of content and tools even more pervasive.

None of these three categories are purely technological issues. They refer to drivers of conflict that predate the current digital era—strategic communications to spread hate through offline media, offline networks of informant-based surveillance, and the polarization effects of conspiracy theories spread by word of mouth, for example. The Rwanda genocide serves as evidence of how these three factors can come together to cause

[2] See for example this course developed by Build Up introducing a framework for digital peacebuilding https://howtobuildup.org/community-learning/courses/digital-peacebuilding-101-introducing-technology-for-peacebuilding/.

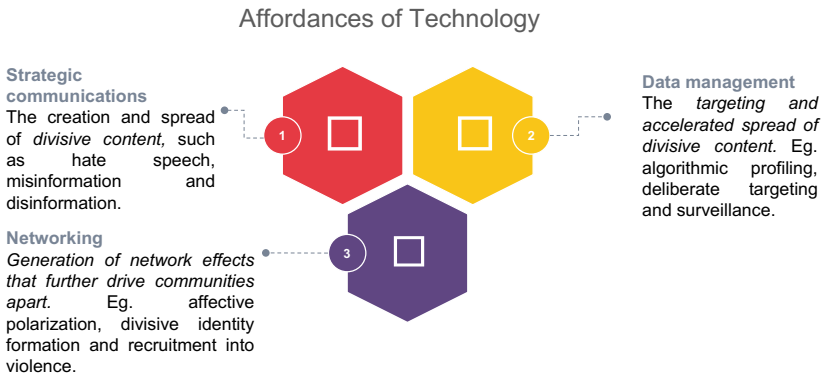

Fig. 9.1 Affordances of technology for conflict

conflict on a major scale, even without the support of modern day technologies. To over-simplify, hate speech was widespread (strategic communications), messages were targeted to specific communities on specific radio stations (data management) and that in turn led to creation of divisive identities whose network effect further fuelled conflict (networking). Thus, digital technologies are not creating *new* drivers of conflict, but rather exacerbating and enabling existing ones. As such, the drivers outlined below are what we call 'socio-technological' issues—societal issues that, when combined with technology, take on new dimensions (Fig. 9.1; Table 9.1).

Strategic Communications

A well-recognised benefit of digital technologies is the ability to share information more widely, at a lower cost and greater speed. Social media in particular, has provided new avenues for sharing information in real time. This is in many ways an opportunity for peacebuilding—from understanding what is happening on the ground as it happens, to diffusing messages that can mitigate or prevent conflict and building campaigns calling for peace.

However, it is also enabling divisive content to be created more easily and spread more rapidly. In particular, there are three key content-types that can serve as conflict drivers.

Table 9.1 Affordances of technology for conflict—examples by category

Strategic communications: digital technologies to *create and spread divisive content*	Hate speech	Using speech, text or images to demean or attack a person as a member of a group
	Misinformation	Spreading incorrect information without the intent to deceive (often as the result of manipulation)
	Disinformation	Creating and spreading incorrect information to intentionally deceive or manipulate others
Data management: digital technologies to *target and accelerate the spread of divisive content*	Algorithmic Profiling	Using large amounts of data to inform personalised recommendation algorithms
	Deliberate targeting	Using digital data to deliver tailored messages with the intent to manipulate specific individuals or groups
	Surveillance	Active collecting of digital data about individuals to exercise control
Networking: digital technologies *to generate network effects that continue to drive communities apart*	Affective polarization	Online behaviours and actions that drive people with different opinions further apart
	Identity polarization	Online behaviours and actions to construct identities that fuel division
	Recruitment	Targeting people with certain online behaviours and actions to recruit them into violent groups and ideological networks

- *Hate speech*: the use of speech, text or images to demean or attack a person as a member of a group
- *Misinformation*: the spreading of incorrect information without the intent to deceive
- *Disinformation*: the creation and spread of incorrect information to intentionally deceive or manipulate others

These phenomena all predate the digital age. However, the way in which they are being shared online—faster and more effectively than before—make them 'digital conflict drivers'.

Hate Speech

'There is no internationally recognised legal definition of hate speech' (UN Strategy and Action Plan on Hate Speech 2019, p. 2). The UN defines it as 'any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor' (UN Strategy and Action Plan on Hate Speech 2019, p. 2). In its community standards, Facebook defines hate speech as 'a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability…we define attack as violent or dehumanising speech, harmful stereotypes, statements of inferiority or calls for exclusion or segregation' (Facebook Community Standards).

The link between hate speech and violence is well explored in research and practice. Studies point to the link between hate speech and violence in Rwanda, for example, even before the current digital age (Yanagizawa 2012). Whilst all forms of hate speech can precede violence, we are particularly interested in hate speech that is relevant to conflict lines. For example, hate speech that references specific subcultures of language regarding a particular party to (or victim of) a conflict. The use of coded language to refer to a specific group within a conflict is a tactic that goes back well beyond the digital era. However, digital technologies have led to an increased ability for hate speech to spread effectively and in a more targeted manner.

In South Sudan, for example, fake news and hate speech spread online and promoted by social media influencers has been used to incite violence. Researchers have shown that social media figures, often based in the diaspora, play an outsize role in influencing events on the ground (Patinkin 2017). Whilst hate speech posted online might not be seen by large swathes of the South Sudanese population due to low internet penetration rates, content shared on Facebook is often then spread through private groups and eventually by word-of-mouth, diffusing its impact well beyond the digital realm. A UN panel of experts report on South Sudan

from November 2016 supports this view, claiming 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, p. 10).

Misinformation

Misinformation can be defined as the spreading of incorrect information without the intent to deceive (often as the result of manipulation). The Covid-19 pandemic led to a flood of health-related misinformation and a concerted international effort to tackle it. In April 2020, the UN Secretary General launched the United Nations Communications Response initiative to combat the spread of mis- and disinformation, in recognition of the particularly potent combination of the pandemic, social media and misinformation.

Just as there are many forms of hate speech, there are many forms of misinformation. For the purposes of this chapter, we are interested in misinformation that has the potential to cause conflict harm. For example, misinformation about an incident of violence or misinformation that peddles false information about a particular community that is relevant to a conflict. Social media platforms themselves make a distinction between misinformation writ large, and misinformation that can contribute to violence or physical harm (Kozlowska 2018).

Misinformation is often widely spread using digital technology. The barriers to entry into the online space are significantly lower in the social media era. Misinformation is often more engaging than verifiable information: it speaks to people's core emotions, it is often punchy with appealing titles, and in general makes for the kind of clickbait that social media algorithms are primed to promote (more on this in later sections).

In addition, once spread, misinformation is particularly hard to combat. It is 'challenging to persuade people with facts once they have adopted a belief or position because of confirmation bias' (The Omidyar Group 2017). As an example, there is some evidence that when Facebook added a flag to show when an article has been disputed by fact-checkers, it in some cases led to increased popularity (Levin 2017).

Examples of misinformation on social media leading to offline violence are numerous. In 2018, rumours about a gang of child abductors spread on WhatsApp in India, sparking a series of mob lynchings that led to at least 17 deaths (Murty 2017). In Sri Lanka, rumours that police had

seized sterilization pills from a Muslim pharmacist led to widespread communal violence (Taub and Fisher 2018). In 2014, a rumour spread on Facebook that a young Buddhist woman had been raped by two Muslim men in Myanmar's second city of Mandalay. In response, a mob formed outside the teashop of the alleged attackers, sparking altercations that led to two deaths (Waheed 2015). The examples are so numerous that the link between misinformation and violence can no longer be disputed.

Disinformation

Disinformation can be defined as the creation and spread of incorrect information to intentionally deceive or manipulate others.

Disinformation differs from misinformation in intent. Whereas misinformation spreads because people share it unwittingly, disinformation is spread as a result of deliberate efforts, often deploying the tactics of deliberate targeting outlined below. As a result of the digital tools available, the opportunities for disinformation to spread are now greater than ever before. Disinformation includes the concept of coordinated inauthentic behaviour—defined by Nathanial Gleicher, head of Facebook security policy as 'groups of pages or people working together to mislead others about who they are or what they are doing' (Gleicher 2018).

Coordinated efforts at disinformation have targeted conflict settings and actors around the world. In Libya, coordinated networks have been used to bolster Khalifa Haftar's Libyan National Army (LNA) (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.

One particularly pernicious form of disinformation is what is sometimes referred to as 'manufactured consensus': using robots to repeat a point of view or fake story online to create the impression that it is mainstream, that many people agree or believe it to be true. This can in turn make it easier to polarize a conversation and harder to find common ground. The structure of social media platforms in turn supports this kind of disinformation, as popularity is often conflated with legitimacy (The Omidyar Group 2017). This tactic has been used in election discourse, to subtly manipulate online discussions in favour of one party—the concept of 'informational measures' (The Omidyar Group 2017).

In Brazil, the use of disinformation during Jair Bolsonaro's election campaign demonstrates the power of disinformation to influence political

outcomes, and to create societal divisions that lead to real-world tensions and violence. During his 2018 run for presidency Bolsonaro's team (and Bolsonaro himself) fuelled the fabricated story that his opponent Fernando Haddad had administered 'gay kits' to indoctrinate Brazil's youth. The story, and many other stories later proven to be false, were disseminated through a vast network of WhatsApp groups by a digitally savvy campaign team (Sidericoudes 2020). The tools of deliberate targeting were put to use, driving Bolsonaro's popularity and fuelling existing social cleavages. Several hate crimes against members of the LGBT community were recorded in the run-up to the elections, with perpetrators making direct links between Bolsonaro's campaign and their attacks (Sullivan 2018).

Data Management

We now turn to a set of data management tools that enable divisive content (i.e. hate speech, misinformation and disinformation) to spread more effectively. Concretely, data can be used to identify groups, to target them with the purpose of manipulation, and to support mass surveillance. These tools serve to further entrench the cleavages opened by the strategic communications affordance.

To some extent, the use of data in this way is not new—it has shaped our lives for a long time. However, the advent of digital technologies in their current form has enabled this to happen on a vastly different scale, and to a much greater level of specificity. This 'Industrial Revolution of Data' has led to such a multiplication of data points that the tools for data management look fundamentally different today than they did even a few years ago.

The use of data is fundamental to the view that technology is not neutral. By gathering masses of data about us—what we do, what we prefer, where we go etc.—technologies can use automated or semi-automated rules (algorithms) to make decisions about what information we are presented, thus nudging and influencing our choices, opinions and behaviours. This alters our experience of reality and can, in a conflict context, alter the dynamics of a given conflict.

There are three main elements of data management that can drive conflict. These are not in themselves issues of conflict. Instead, they serve as crucial tools that, when combined with divisive content, can spread division and sow the seeds of violence.

- *Algorithmic profiling*: The way that algorithms are structured provides us with polarizing content
- *Deliberate targeting*: Enabling actors to deliberately increase polarization through profiling
- *Surveillance*: Collecting data and using that data to exert control.

To distinguish between these three elements, consider a scale getting deeper the further you go from profiling to targeting to surveillance. Profiling is about how algorithms are structured to serve up things that polarize us. Targeting is about how we can deliberately increase that polarization by using profiling. Surveillance is about going out to deliberately collect data and then using it to exert control.

Algorithmic Profiling

In 2020, the average internet user created 2.5 quintillion bytes every day (Bulao 2021). Using mobile phones to browse the internet, update social media profiles, shop, navigate and follow the news, individuals across the world are leaving behind them a vast trail of data points. These individual data points are in many cases protected as users and regulators become more aware of privacy rights, although even where individually identifiable data is protected, aggregate use of the same data may be permitted.

The power of this data lies in creating profiles of people with similar characteristics. When coupled with powerful algorithms (recommender systems that determine what we will like based on information from our user profile), these profiles inform many aspects of our lives—from what we see when we browse the internet and who we befriend on social media to how easy it is for us to find a job (Tisne 2018).

This algorithmic profiling is particularly concerning on social media and search platforms, where algorithms target us with information and narratives that we are most likely to agree with. They expose us only to certain people or experiences. If we show some tendencies towards a certain opinion, the algorithm will entrench that tendency by showing us more of that type of content. 'You watch one video that's lightly critical of feminism…and YouTube's algorithm leads you down a rabbit hole of videos that grow increasingly misogynistic, never urging you to stop or change course' (Wood 2019). Algorithms are providing positive feedback loops to division, fuelling additional polarization, even where the user is not actively seeking it.

This is problematic for two reasons. First, it creates different realities depending on who we are. This is concerning from a conflict perspective because it further drives people apart and reduces any sense of common experience. Second, when coupled with a focus on user engagement at all costs it exposes people to more extreme positions. This is problematic from a conflict perspective because it makes it harder to find common ground or third poles for dialogue.

Social media algorithms are designed to promote engagement, and particularly 'affective engagement': creating emotional reactions to content based on flashes of positive or negative feeling. Social media platforms compete in an 'attention economy', and are therefore focused on winning as much audience engagement as they can (Bhargava and Velasquez 2020). As a result, more extreme, violent or polarizing content tends to drive more engagement, so algorithms amplify divisive content over more neutral content. In this way, they funnel users towards more extreme content. As explored in a study by William J. Brady et al. (2017), Tweets using moral and emotional language receive a 20% boost for every moral and emotional keyword used.

This algorithmic focus on affective engagement not only ensures that more emotional and divisive posts spread more widely on social media, but in some cases, they can also gain in perceived legitimacy as a result of this sharing. These posts then take on greater significance in the wealth of available information, in turn influencing the worldview of those who see them.

This means we have two results: algorithmic profiling divides people according to their preferences and it turns those preferences more divisive and extreme.

Numerous examples highlight the real-world impact of this algorithmic profiling on conflict—often through the radicalisation of individual viewpoints towards the extremes. Caleb Cain (2019) has openly documented how he 'fell down the alt-right rabbit hole', falling prey to the cycle of YouTube recommendations. Through YouTube's 'Up Next' recommendation, Cain—who started out as a 'liberal college drop-out'—found himself being drawn closer to a radicalised ideology, eventually ending up deep in an alt-right community. His journey highlights how the YouTube recommendation algorithm makes assumptions about individuals that in turn can push them towards more extreme content, driving conflict.

Facebook uses algorithmic profiling to help users expand their individual networks through the 'suggested friends' feature. This feature uses

algorithmic profiling to help bring like-minded individuals together. But in 2018, research emerged showing that algorithmic profiling has helped terrorists build networks of like-minded individuals (Ratner 2018). By bringing people of like-mind together, algorithmic profiling inadvertently served to bolster extremist networks.

Deliberate Targeting

Whilst algorithmic profiling serves to provide users with divisive content through automated recommendation formulae, the concept of deliberate targeting takes this to the next level.

Digital technologies, and social media and search platforms in particular, make it possible to deliver tailored messages to individuals, or to groups of people based on certain characteristics. This targeting has been used by conflict actors both to undermine individuals whose views are opposing theirs, and to manipulate specific groups with the intent of creating divisions.

Doxxing—the deliberate leaking of personal information about an individual online for harassment or negative intent—can be used to undermine the efforts of activists or those critical of a regime. Doxxing begins with a search for online clues that can then be used to reveal private information such as passwords, enabling the perpetrators to build detailed profiles of their targets. In a conflict setting, doxxing can be used to undermine opponents or to discourage peace efforts by targeting prominent peace activists.

In Hong Kong, the use of doxxing has targeted individuals on all sides of the conflict between government and protesters. Student protesters have had their personal information revealed, many on a site called HKLeaks, which 'targets activists, journalists, social workers and even media magnates' (Borak 2019). Borak explains how private messaging channels such as Telegram have also been used to spread personal information—of both pro and anti-Beijing individuals—in an attempt to intimidate individuals and prevent them from further engaging in protests. There are increasing fears that this tactic leads to real world violence, as some victims of doxxing have already faced attacks. Several protesters have received threatening calls following the leak of their personal information, whilst other offline events have been linked to doxing (MENAFN 2020).

Whereas individual victims of doxxing will often know they have been targeted, subtler forms of targeting exist that are perhaps more concerning, largely because those who have been targeted are rarely

cognisant of the targeting. As with data aggregation, this tactic is not unique to the digital era. However, the ability to target with an ever-increasing degree of specificity, when coupled with the data discussed above, makes the current issue uniquely challenging.

Highly targeted content is increasingly being deployed as a tactic by divisive actors intent on polarizing conversations. Using the data aggregation discussed above and the sophisticated advertising tools provided by social media platforms and search platforms, individuals are able to serve particular groups with particular content. This is of particular concern in conflict settings, where it becomes possible to target groups along conflict lines with divisive content, further driving opposing groups apart.

Group targeting uses certain characteristics to build audiences that in turn can be sent specific messages. Facebook Ads audience creation setting, for example, allows the administrator of any Facebook page to design sophisticated audiences based on demographic and geographic information. Audiences can be built based on criteria such as location, age, gender, language, as well as interests and connections. Administrators can then run particular Ads to one specific group.

Once an audience is made, A/B testing enables content creators to test different messages to maximise their engagement. This tactic enables advertisers to refine their techniques by showing them which content is working best for which audiences. All of this can be done on a very limited budget.

One step up from group targeting is the ability to target specific individuals. 'Sniper targeting' enables pernicious actors to hone in on individuals they want to receive a particular narrative. The use of sniper targeting by a disillusioned Mormon to convince his wife to abandon the church is well documented—and was so successful that the user deployed the tactic on several other members of the Mormon community (Faddoul et al. 2019).

Targeting is a marketing tactic—we are not disputing it. However, the ability to micro-target can serve to drive conflict in two core ways. First, it allows actors to foster division by targeting groups along conflict lines. It can be used to create parallel narratives that in turn further drive people apart. Second, it is often used in parallel with problematic content, such as misinformation along conflict lines.

Perhaps the most well-known example of deliberate targeting causing real-world division is the Cambridge Analytica scandal, which used data from 50 million Facebook users to target specific groups with unique

advertisements in the run-up to the 2016 US election. Data collection was done via a personality quiz which enabled Cambridge Analytica to gather large amounts of data on individual characteristics. This, coupled with vast amounts of Facebook data and electoral register records, enabled a detailed profiling of individuals which was in turn used to deliberately target specific groups with specific messages through Facebook Ads. Whilst measures have been taken to limit the particular data collection method utilised by Cambridge Analytica, micro-targeting remains a highly accessible tool for actors seeking to promote certain ideas to certain communities.

Surveillance

Third, data technologies make it possible for certain actors—especially governments—to actively collect digital data about individuals at a large scale. This kind of surveillance data can be used to exercise control, especially in conflict situations.

Surveillance to control is not a new tactic, but the same digital exhaust that is used to create profiles for marketing can be used to track and control individuals. This can be done by asking technology companies to hand over certain information (usually a prerogative of governments), by deploying hacking tactics to access this data from the companies or from individuals (e.g. through malware), or by directly collecting individually identifiable data and processing it with artificial intelligence (e.g. AI-powered CCTV networks).

Where algorithmic profiling affects most of us and deliberate targeting is available to many conflict actors, the use of surveillance data to control groups in ways that contribute to conflict is limited to a smaller set of actors who have the capacity to command sufficient data through one of the three techniques above. These actors may be fewer, but their impact in conflict contexts can be deeper. In Venezuela, the Maduro government has used the Homeland ID card as a mechanism to exert control over citizens. The cards link users' data to a government database and connect holders to social welfare platforms through digital QR codes, enabling the government to keep tabs on its citizens (Puyosa 2019). Puyosa explains how in the 2018 presidential election, the government linked food distribution to individuals passing through pro-Chavist kiosks, demonstrating the power of digital surveillance to influence citizens.

In response to COVID-19, many governments have rolled out track and trace systems that rely on location surveillance data. The Electronic

Frontier Foundation has sounded the alarm about these measures being rolled out too quickly and with little regard for digital rights (Schwartz and Crocker 2020). In Israel, for example, the move by the government to use geolocation data collected by cellphone providers to track the spread of the virus was hotly contested as it could open the door to tracking individuals across conflict lines (Halbfinger et al. 2020).

Not all surveillance is conducted by governments though. In South Africa, a number of private companies (most notably Vumacam) are driving the roll-out of smart CCTV systems across most major cities. These AI-powered systems scrutinise peoples' demographics and movement for a pre-coded set of unusual behaviours that only thinly disguise a racial bias, exacerbating post-apartheid injustice and tensions (Kwet 2019).

Dialogue and Networking

Strategic communications refers to the divisive *content* that technology enables the spread of. Data management refers to a set of *tools* that enable that content to further divide. The networking affordance in turn creates a set of *enabling conditions* that further enhance division.

To understand how networking is used to increase divisions, we explore three different aspects.

- *Affective polarization*: the way that technology serves to drive people apart
- *Construction of identities*: the importance of technology in constructing certain identities
- *Recruitment*: the use of technology to recruit individuals into violent actions.

Affective Polarization

Polarization refers to a set of behaviours and actions, intended and unintended, that drive people with different perspectives further and further apart. Affective polarization is division that in turn leads to relational group hate. Whilst some researchers posit that polarization is not happening *because of* technology (Laurenson 2019), there are some online behaviours and actions that drive people with different opinions

further apart—even if that is not their intention. This kind of polarization is driven by certain features of online technologies.

Take for example, the structure of social media. We know from the above sections that misinformation and divisive content has increased reach. We know that the tools of targeting and algorithmic profiling can make that reach even more specific. In addition, certain features of social media serve to simplify narratives and further polarize communities. Conversations on digital platforms are short, immediate and publicly recorded, making quickly identifying with positions more attractive than slowly parsing out common needs.

These features serve to make the online environment one of polarized positions, where constructive disagreement and debate is rarely seen. You might think that people would flee such a rarified environment, but we know they don't. This is partly explained by what we explored in the first section: polarizing content is more appealing to the human brain, triggering neurological responses that satisfy a need to belong and build human capital, and in this sense is somewhat addictive. Social media platforms know this, and they have built notification and nudge features to build on this addiction—with each ping comes another dopamine hit.

Affective polarization matters for conflict. With polarization, comes the absorption of neutral actors into increasingly more rigid and extreme positions taken in opposition to other factions. In turn, polarization supports the strengthening of convictions in different factions, making it less likely for someone to break from their personal value system. Finally, polarization can result in distorted perceptions and simplified stereotypes along with diminished trust or agreement with other factions over basic facts and realities. The combination of these factors contributes to limited opportunities or desire for shared dialogue. Well-established models of conflict escalation signal that these constitute warning flags for future violent confrontations. Indeed, 'conflict theorists pay attention to polarization because increased polarization is a warning sign for armed conflict' (Laurenson 2019, p. 3).

Affective polarization serves to sharpen the impact of the content and tools outlined above. When misinformation goes viral, it does so in an already polarized environment. When deliberate targeting is used to serve a particular narrative to a community, that community is already being pushed further away from its counterparts. In essence, spreading hate speech or engaging in deliberate targeting is adding fuel to an existing fire. As such, looking at divisive content or micro-targeting as

isolated instances with impact on certain individuals is to miss the broader, systemic issue that results from the network effects of a polarized online environment.

Online Identity Construction

The tools and content outlined above can serve as conflict drivers regardless of the online environment. However, just as polarization can serve to make those issues more dangerous, so too does the impact of these technologies on individual identity formation. Research has shown that social media alters our incentives and affects how we construct discourse—as a result it impacts how our collective and individual identities are shaped and expressed. The role that social media has in constructing identity can and has been exploited by conflict actors. Actors can use the tools of deliberate targeting to spread disinformation with the intention of constructing identities that fuel division. The internet is rife with digital tribes that oppose other digital tribes.

This doesn't affect as many people as affective polarization, but it can be a powerful force in divided societies. Identity formation is well documented as a driver of conflict in the offline space.[3] As that process is increasingly playing out online, understanding the link between digital technologies and divisive identity formation is crucial (although still underexplored). Recent research by Build Up gathered examples of social media conversations that impact identity formation in conflict settings. They found that the most common way to express divisive identities online is to dismiss an opinion or position through mockery. Fake news was a key theme in the expression of divisive identities, often to accuse a group of a behaviour or associate them with a negative identity marker. In addition, the research found that the way conversations unfold on social media can contribute to deepening divisions about identities, for example through escalation in comment threads that lead to extreme position.[4]

Recruitment

Finally, some actors use digital tools to recruit people into violent groups and ideological networks. Hallmarks of extremist recruitment include

[3] See for example https://www.beyondintractability.org/userguide/identity-conflicts.

[4] See Build Up present the results of their research at the Stockholm Forum 2020 on 'Online Identity Formation: A Growing Challenge to Peace' here https://www.youtube.com/watch?v=dbCwaM20kXQ.

the sharing of videos depicting violence (such as the livestream of the Christchurch mosque shooting in 2019), hate speech and targeted online messages. Sniper targeting, for example, could serve to identify specific individuals at risk of radicalisation and enable groups to target them with recruitment materials tailored to their specific identity. This, coupled with the affective polarization and identity formation outlined above, can make for a potent cocktail of division.

This is a more sustained and aggressive approach than the construction of conflict identities online, but it usually flows from it.

Part 3: Peacebuilding Responses

Part 2 outlines digital conflict drivers; Part 3 looks into what peacebuilding practitioners could do about them. Most responses to date have focused on one aspect of these challenges: the most egregious or evident forms of digital harm, including recruitment, hate speech and overt targeting. Few attempts have been made to lay out a comprehensive framework for peacebuilding responses to deeper, or less evident, digital conflict drivers.

The pyramid below attempts to set out such a framework for peacebuilding responses to digital conflict drivers. Each layer of this pyramid requires different approaches, and no single actor is well equipped to address the full spectrum of issues. However, this framework can help peacebuilding actors situate their work, recognise where they add value and coordinate with other actors. It is our hope that this could lead to the beginning of a more holistic view of the emerging field of 'digital peacebuilding' (Fig. 9.2).

Level 1: The Signal

At the highest level, we have the 'signals'—things that we can easily see when looking at the digital environment. These include hate speech, surveillance, recruitment and the overt targeting of individuals e.g. sniper targeting or doxxing. These activities often point to deeper issues that are driving conflict. Hate speech spreading on social media, for example, is often a signal of deeper conflict issues.

To date, much of the discourse on these signals has focused on their removal—hate speech reporting, content moderation or restrictions on overt targeting. These approaches are important, but they only go so far.

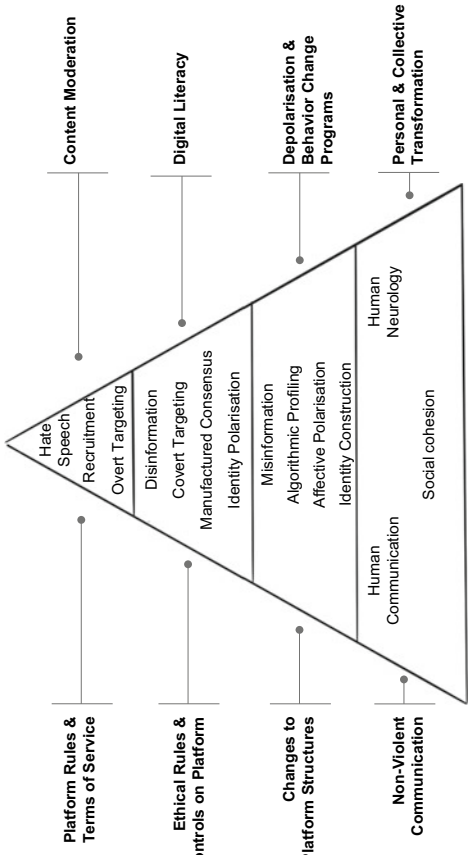

Fig. 9.2 Pyramid of digital conflict drivers and respective peacebuilding responses

First, by 'burying the signal', we risk ignoring the underlying issues. To remove hate speech after it has occurred, but not to address the deeper issues is equivalent to addressing the symptoms but not the disease. In addition, moderating hate speech on one platform simply results in that speech going elsewhere—and often to places where it is harder to find (but no less powerful in reaching its intended audience).

Peacebuilding Responses

Hate Speech Monitoring and Content Moderation

The monitoring of hate speech is often done by civil society actors or NGOs, largely as a way to better understand the issue rather than to resolve it directly. The PeaceTech Lab, for example, has developed a set of 'hate speech lexicon' in order to define problematic content in countries such as Yemen, Sudan and Kenya.[5] These lexicon and other efforts to monitor and report hate speech can serve to inform other peacebuilding responses to these issues.

The response of social media platforms to hate speech has been focused on content moderation. To detect and remove hate speech, platforms rely on a combination of Artificial Intelligence and user reports. These user reports often come from civil society actors who proactively report content to platforms such as Facebook. To support increased detection of hate speech, Facebook has been working to educate users on their community standards—the rules of engagement with the platform that explicitly prohibit hate speech. In Myanmar, where Facebook has faced significant backlash against its failure to tackle hate speech, Facebook published its community standards in Burmese and hired additional Burmese speaking monitors, in order to increase its detection of hate speech. To date, the effectiveness of content moderation efforts is limited—hate speech continues to be present on social media platforms. In addition, content moderation does not address any of the underlying causes that enable hate speech to thrive online.

Countering Activities

Responses that seek to counter, rather than remove, the effects of hate speech recognise the limitations of content moderation and reporting. There is a wide body of work on effective tools to counter hate speech

[5] See for example, https://www.peacetechlab.org/toolbox-lexicons.

online. These methods emphasise, for example, shared identity, demonstrating intergroup friendships, whilst cautioning around employing empathy-only approaches. Effective responses often involve the community in the development of counter-messages ensuring that narratives unite rather than divide the public square. Countering methods have also been used to interrupt recruitment efforts into extremist groups.

#IAmHere is a network of tens of thousands of online volunteers fighting hate speech on Facebook (Bateman 2019). Volunteers scan Facebook for conversations happening on popular pages, often run by mainstream media organisations, which are overwhelmed with racist, misogynistic or homophobic comments. Volunteers don't attempt to change the minds of people posting hate or argue directly with extremists. Instead they collectively inject discussions with facts and well-argued reasonable viewpoints. The idea is to provide balance so that other social media users see that there are alternative perspectives beyond the ones offering up hate and division.

The Institute for Strategic Dialogue's *Counter Conversations* programme seeks to counter recruitment into radical groups (Davey et al. 2018). ISD, a global think tank dedicated to countering extremism, identified that extremist groups deploy a clear strategy for radicalising and recruiting new supporters online: marketing their ideas through the spread of propaganda and then engaging interested individuals in direct, private messaging to recruit new members to their causes. The Counter Conversations programme identified individuals who were demonstrating signs of radicalisation on Facebook, and engaged these individuals in direct, personalised and private 'counter-conversations' on Facebook Messenger for the purpose of de-radicalisation from extremist ideology and disengagement from extremist movements.

Education on Privacy

Whilst combating government-led surveillance is a challenge, many civil society actors have taken the approach of informing citizens about their digital safety, encouraging them to enhance their privacy settings.

SalamaTech is a project of the SecDev Foundation, a Canadian think tank that works at the cross-roads of conflict, development and new technology.[6] Since 2012 SalamaTech has helped Syrian peacebuilders

[6] You can view their work at https://en.salamatech.org/.

stay safe online so they can make their voices heard. Syrian civil society actors are increasingly targeted through cyberspace by a range of actors. Digital threats manifest across multiple and distributed channels, through targeted attacks, profiling of personnel and supports, and theft of sensitive information.

SalamaTech assists Syrians who have had their accounts hacked. They protect Syrian civil society organisations with Digital Safety Audits, which build the capacity of CSOs to protect their data and use the internet safely. They have a network of on-the-ground Digital Technology First Responders who provide in situ training to protect Syrian civil society organisations. By providing these protections, SalamaTech ensures that Syrian CSOs can continue their work.

Level 2: Below the Surface

Just below the surface lie the second set of digital conflict drivers—issues that are somewhat harder to find than those at the top, but that still point to deeper drivers below. In this category we place disinformation, covert targeting, manufactured consensus and the construction of identities online. These are issues that the end user may not be aware that they are exposed to (for example group targeting), but that can be identified with some level of awareness. Although less familiar than those at the surface level, these challenges have received increasing attention recently. The Covid-19 pandemic saw an increased prevalence of disinformation, leading the UN to declare an 'infodemic', in turn prompting a growing number of conversations about how to tackle it (UNODC 2020). At the same time, discussions around covert targeting, manufactured consensus and identity construction have been recently making the news and entering popular discourse.

Peacebuilding Responses

Debunking Disinformation

Efforts to combat disinformation take many forms, from researchers developing ways to detect disinformation, to those working to debunk disinformation once it emerges.

In Lithuania, a group of citizen volunteers known as 'elves' tackle Russian-driven disinformation. They work to identify and address disinformation through a combination of tactics designed to mitigate the divisive potential of these efforts (Peel 2019).

Social media platforms seek to tackle disinformation through the identification and removal of coordinated campaigns. Policies on inauthenticity and on coordinated inauthentic behaviour seek to remove deliberate attempts to deceive or manipulate the debate through inauthentic means. Facebook's policy on inauthentic behaviour bans users from 'artificially boosting the popularity of content' (Facebook). However, their policies on disinformation remain opaque, with several civil society actors criticising the platforms for failing to share the criteria on which they decide what counts as misinformation. Others criticise platforms (particularly Facebook) for their restrictive sharing of data, preventing researchers from getting a granular analysis of the problem.

Social Cohesion Campaigns

Supporting the construction of common identities and combating the creation of divisive identities is a pillar of traditional peacebuilding. Several peacebuilding organisations are now taking that work into the online space. For example, online campaigns that call for the construction of common identities. The *Peace Factory* seeks to forge common identities and tackle divisive identity positioning in the Middle East through campaigns such as 'Israel loves Iran'.[7] In Myanmar, a local civil society organisation led a Facebook campaign seeking to bring young people together around a common identity (the names of the organisation and campaign have been removed to protect the identity of those involved).

Integrating Social Media into Peace Agreements

To date, most peacebuilding responses to disinformation have sought to address the problem 'downstream' (i.e. tackling the distribution and consumption of disinformation). Fewer responses have sought to intervene 'upstream' (i.e. preventing the production of disinformation). In response, some mediation actors have begun exploring the concept of 'social media peace agreements', or the integration of social media clauses into existing peace agreements and dialogue processes. This work aims to discuss social media activity directly with conflict parties, and facilitate agreements in which parties agree to exercise restraint in the digital space. Whilst still a relatively new area of intervention, the Centre for

[7] You can view their campaigns at https://www.facebook.com/the.Peace.Factory/.

Humanitarian Dialogue (HD) is leading efforts to forge such agreements, including around elections. In Indonesia's 2020 local elections, for example, HD helped facilitate a social media code of conduct in an attempt to restrain the production of disinformation around the election.

Level 3: Across the Board

Below these challenges, lie the issues of misinformation, algorithmic profiling and affective polarization. These more deep-seated issues require a different set of approaches from peacebuilding actors. In addition, they are less often discussed and addressed by peacebuilding actors, and still lack mainstream attention in the conflict field.

Peacebuilding Responses

Combatting Misinformation

Social media platforms have taken several steps to combat misinformation—from labelling content as misinformation, to informing people that content has been debunked by fact checkers before they share it, to removing misinformation that can contribute to violence or physical harm. Facebook has fact-checking partnerships with multiple civil society organisations that help them spot and tackle misinformation on their platform. WhatsApp introduced measures to limit the sharing of messages to more than five people in response to growing problems of misinformation on its platform. Twitter adds labels to provide warning messages and context on Tweets containing misinformation. In response to the Covid-19 pandemic and in the midst of growing public awareness of the issue, platforms expanded their misinformation policies.

Digital Literacy

In response to the increasing challenge of misinformation, there has been an uptick in digital literacy efforts. These have been led by civil society organisations, local and international NGOs, UN agencies and social media platforms themselves.

The *Myanmar ICT for Development Organisation (MIDO)* works to promote technology for social change in Myanmar and to promote media and digital literacy through a range of programmes. They run a Facebook page that incorporates a Messenger chatbot to promote media literacy in Myanmar. MIDO's page and chatbot provide a fact-checking function where people can report in rumours and hear back within hours whether

the MIDO team can verify the rumour, serving as a novel way to debunk misinformation. A separate chatbot also provides media literacy e-learning content. In a context where low media literacy is closely tied to intercommunal conflict and misinformation contributes to increasing polarization and division, MIDO's work helps minimise the impact of fake news and misinformation that can incite intercommunal violence.

Policy Responses

Many governments have introduced policies that address aspects of misinformation: Section 230 of the Communications Decency Act in the USA and the Digital Services Act in the EU are two such policies whose reach and applicability are currently much debated. Several advocacy bodies are calling for stronger regulation to stop data being misused. Some of these call for greater individual ownership of data, others for more transparency on how data is used.

Martin Tisne (2018) calls for *A Bill of Data Rights* that would give people rights to decide how their data is used. Tisne argues that existing discussions based on the idea of 'data ownership' are flawed as they ignore the broader problem of how data is used in the aggregate. His proposed bill of data rights could start from the following principles:

- The right of people to be secure against unreasonable surveillance shall not be violated.
- No person shall have his or her behaviour surreptitiously manipulated.
- No person shall be unfairly discriminated against on the basis of data.

Depolarization Efforts

Build Up's *The Commons* project seeks to tackle depolarization on social media in the United States (Build Up 2019). The project identifies people engaged in political conversations on Twitter and Facebook in the USA, analyses what kinds of behaviours may denote a person is exposed to polarizing narratives or dynamics, and targets people with these characteristics with automated messages that invite them into a conversation about bridging divides.

If they respond, one of Build Up's trained dialogue facilitators has a conversation with them on the platform (Twitter and Facebook). These facilitated conversations seek to help people understand and make

different choices in their interactions, online and offline, particularly around political differences, and offer skills and resources to promote constructive conversations, listening and respect.

Redesigning Algorithms

Several researchers are proposing the redesign of social media platform algorithms, to counter the divisions fostered by algorithmic profiling. Amongst others, Helberger et al. (2016) explore what a 'diversity-sensitive design' would look like when applied to algorithms. They posit a redesign of algorithms that would actively encourage diverse exposure to information, breaking down filter bubbles. Laurenson (2019) refers to research that distinguishes between 'connection-promoting' vs 'non-connection promoting' social media use. She explores the possibility of designing social media platforms in a way that could model less polarizing interactions on social media. Rose-Stockwell (2018) in turn suggests the retraining of social media algorithms to refine the concept of 'meaningful content'. He suggests excluding content that is categorised as 'outraged, toxic and regrettable' from algorithms' definition of meaningful.

Countering the Algorithm

The *Redirect Method* is a project of Moonshot CVE, a technology company that works to counter violent extremism.[8] The project analysed how people searching for certain words or phrases on Google find ISIS videos on YouTube, aided by algorithmic profiling.

To counter this, Moonshot CVE bought Adwords on Google—these are advertisements that can be targeted at people who search for certain words or phrases. They used these advertisements to ensure that when people searched for words or phrases that would have previously led them to ISIS propaganda videos, they instead saw videos debunking ISIS recruitment themes. This open methodology was developed from interviews with ISIS defectors. It also respects users' privacy and can be deployed to tackle other types of violent recruiting discourses online.

[8] You can find out more about this at https://redirectmethod.org/.

Level 4: The Roots

At the very bottom of the pyramid lie the most intractable parts of this digital conflict context. These issues have received the least amount of attention to date, and require research that goes well beyond the limits of this chapter. However, there is growing evidence that digital technologies are not only affecting our context, but our brains as well. Researchers have explored links between social media addiction and dopamine levels (Parkin 2018). An open question remains about how much technologies are, for example, shifting our incentives, or altering our decision-making processes.

Whilst these are open questions, it is clear that digital conflict drivers touch on some of the deepest roots of the human condition—our mode of communication, our neurology and, ultimately, how we live together. Few peacebuilding approaches have yet sought to address these digital conflict drivers directly. However, we believe that the tools that have been used by peacebuilders to bring about personal and collective transformation in the offline realm may be needed in the digital realm. Ultimately, approaches such as non-violent communication may help to shift behaviours in a way that allow us to build peace in today's new socio-technological context.

Conclusion

In our increasingly connected world, the distinction between online and offline elements of a conflict is no longer clear. As technology has become deeply intertwined with our experience of the world around us, it has begun to shape the structures of power that bind us as communities and define us as individuals. It no longer makes sense for peacebuilders to view technology as separate from the conflict context—as either a tool for positive change or as a weapon for fuelling war. Instead, we need to understand technology as integral to a context, and dig deeper into the dynamics of socio-technological conflict. The above framework for categorising digital conflict drivers can advance this goal. Peacebuilders can then begin to situate their interventions in the pyramid of responses outlined above, connecting their work more deliberately to the complex links between technology and conflict. In doing so, it is our hope that peacebuilding as a field won't just tackle the surface level problems in the

use of digital technology, but take on the most challenging aspects of how conflict is evolving and re-emerging in the digital age.

References

Bateman, J, 9 June 2019, '#IAmHere': The People Trying to Make Facebook a Nicer Place', *BBC Trending*, Accessed 23 February 2021 https://www.bbc.com/news/blogs-trending-48462190

Bhargava, V. R. and Velasquez, M., 2020, 'Ethics of the Attention Economy: The Problem of Social Media Addiction', *Business Ethics Quarterly*. Cambridge University Press, pp. 1–39.

Build Up, 12 December 2019, 'Scaling the Commons: An intervention to depolarize Political Conversations on Twitter and Facebook in the USA', *Medium*, Accessed 23 February 2021, https://howtobuildup.medium.com/scaling-the-commons-969b15c98012

Borak, M, 10 October 2019, 'Doxxing Has Become a Powerful Weapon in the Hong Kong Protests', *Abacus*, Accessed 19 February 2021, https://www.scmp.com/abacus/culture/article/3032268/doxxing-has-become-powerful-weapon-hong-kong-protests

Brady, W, Wills, J, Jost, J, Tucker, J and Van Bavel, J, 2017 'Emotion Shapes the Diffusion of Moralized Content in Social Networks' *PNAS*, Accessed 19 February 2021, https://drive.google.com/file/d/0B3NfwG0qz3qOaUl5bVZpQWRWc2c/view

Bulao, J, 22 January 2021, 'How Much Data Is Created Every Day in 2020', *TechJury*, Accessed 19 February 2021, https://techjury.net/blog/how-much-data-is-created-every-day/#gref

Cain, C, 8 June 2019, 'The Making of a YouTube Radical', *The New York Times*, Accessed 16 February 2021, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

Davey, J, Birdwell J and Skellet R, 2018, 'Counter Conversations: A Model for Direct Engagement with Individuals Showing Signs of Radicalization Online', *ISD*, Accessed 23 February 2021, https://www.isdglobal.org/wp-content/uploads/2018/03/Counter-Conversations_FINAL.pdf

Facebook, Community Standards, *Hate Speech*, Accessed 16 February 2021 https://www.facebook.com/communitystandards/hate_speech

Facebook, Community Standards, *Inauthentic Behaviour*, Accessed 23 February 2021, https://www.facebook.com/communitystandards/inauthentic_behavior

Faddoul, M, Kapuria, R and Lin, L, 10 May 2019, 'Sniper Ad Targeting', Accessed 19 February 2021, https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/sniper_ad_targeting_final_report.pdf

Gleicher, N, 6 December 2018, 'Coordinated Inauthentic Behavior Explained' *Facebook Newsroom*, Accessed 17 February 2021, https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

Grossman S, H K, DiResta, R, Kheradpir, T and Miller C, April 2 2020, 'Blame It on Iran, Qatar, and Turkey: An Analysis of a Twitter and Facebook Operation Linked to Egypt, the UAE, and Saudi Arabia', *Stanford Internet Observatory: Cyber Policy Center*, Accessed 17 February 2021, https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/20200402_blame_it_on_iran_qatar_and_turkey_v2_0.pdf

Halbfinger, D, Kershner, I and Bergman, R, 16 March 2020, 'To Track Coronavirus, Israel Moves to Tap Secret Trove of Cellphone Data', *The New York Times*, Accessed 23 February 2021, https://www.nytimes.com/2020/03/16/world/middleeast/israel-coronavirus-cellphone-tracking.html

Helberger, N, Karppinen, K and D'Acunto, L, 2016, 'Exposure Diversity as a Design Principle for Recommender Systems', *Information, Communication & Society*, 21(2), 191–207, Accessed 19 February 2021, https://doi.org/10.1080/1369118X.2016.1271900

Kozlowska, H, 19 July 2018, 'Facebook Is Actually Going to Start Removing Fake News—Or Some of It', *Quartz*, Accessed 16 February 2021, https://qz.com/1331476/facebook-will-start-removing-fake-news-that-could-cause-harm/

Kwet, M, 22 November 2019, 'Smart CCTV Networks Are Driving an AI-Powered Apartheid in South Africa', *Vice*, Accessed 19 February 2021, https://www.vice.com/en/article/pa7nek/smart-cctv-networks-are-driving-an-ai-powered-apartheid-in-south-africa

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

Levin, S, 16 May 2017, 'Facebook Promised to Tackle Fake News. But the Evidence Shows It's Not Working', *The Guardian*, Accessed 16 February 2021, https://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-working

MENAFN, 18 September 2020, 'Anonymous Site Ramps Up 'Doxxing' Campaign Against HK Activists', *MENAFN*, Accessed 23 February 2021, https://menafn.com/1100816816/Anonymous-site-ramps-up-doxxing-campaign-against-HK-activists

Murty, B V, 22 May 2017, 'Jharkhand Lynching: When a WhatsApp Message Turned Tribals into Killer Mobs', *Hindustan Times*, Accessed 16 February 2021, https://www.hindustantimes.com/india-news/a-whatsapp-message-claimed-nine-lives-in-jharkhand-in-a-week/story-xZsIlwFawf82o5WTs8nhVL.html

Parkin, S, 4 March 2018, 'Has Dopamine Got Us Hooked on Tech?', *The Observer*, Accessed 19 February 2021, https://www.theguardian.com/technology/2018/mar/04/has-dopamine-got-us-hooked-on-tech-facebook-apps-addiction

Patinkin, J, 15 January 2017, 'How to Use Facebook and Fake News to Get People to Murder Each Other', *Buzzfeed News*, Accessed 16 February 2021, https://www.buzzfeednews.com/article/jasonpatinkin/how-to-get-people-to-murder-each-other-through-fake-news-and

Peel, M, 4 February 2019, 'Fake News: How Lithuania's 'Elves' Take on Russian Trolls', *Financial Times*, Accessed 23 February 2021, https://www.ft.com/content/b3701b12-2544-11e9-b329-c7e6ceb5ffdf

Puig Larrauri, H. and Kahl, A, 2013, 'Technology for Peacebuilding', *Stability: International Journal of Security and Development*, 2(3), p.Art. 61, Accessed 15 February 2021, https://doi.org/10.5334/sta.cv

Puyosa, I, November 2019, 'Venezuela's 21st Century Authoritarianism in the Digital Sphere', *Toda Peace Institute. Policy Brief N°62*, Accessed 23 February 2021, https://toda.org/assets/files/resources/policy-briefs/t-pb-62_iria-puyosa_venezuelas-21st-century-authoritarianism.pdf

Ratner, P, 11 May 2018, 'How Facebook Helps Members of ISIS and Other Extremists Find Friends', *Big Think*, Accessed 19 February 2021, https://bigthink.com/paul-ratner/how-facebook-algorithms-helped-isis-and-continue-to-aid-extremist-groups

Rose-Stockwell, T, 30 April 2018, 'Facebook's Problems Can Be Solved with Design', *Quartz*, Accessed 19 February 2021, https://qz.com/1264547/facebooks-problems-can-be-solved-with-design/

Schwartz, A and Crocker, A, 23 March 2020, 'Governments Haven't Shown Location Surveillance Would Help Contain Covid-19', *Electronic Frontier Foundation*, Accessed 18 February 2021, https://www.eff.org/deeplinks/2020/03/governments-havent-shown-location-surveillance-would-help-contain-covid-19

Sidericoudes, S, 9 March 2020, 'The Power of Bolsonaro's Message in his Campaign', *Diggit Magazine*, Accessed 19 February 2021, https://www.diggitmagazine.com/articles/bolsonaro-message-campaign

Stanford Internet Observatory, 15 December 2020, 'Stoking Conflict by Keystroke', *Stanford Internet Observatory*, Accessed 19 February 2021, https://cyber.fsi.stanford.edu/io/news/africa-takedown-december-2020

Sullivan, Z, 29 October 2018, 'LGBTQ Brazilians on Edge After Self-Described 'Homophobic' Lawmaker Elected President', *NBC News*, Accessed 19 February 2021, https://www.nbcnews.com/feature/nbc-out/lgbtq-brazilians-edge-after-self-described-homophobic-lawmaker-elected-president-n925726

Taub, A and Fisher, M, 21 April 2018, 'Where Countries Are Tinderboxes and Facebook Is a Match', *New York Times*, Accessed 19 February 2021, https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html?auth=login-email&login=email

The Omidyar Group, 1 October 2017, 'Is Social Media a Threat to Democracy', Accessed 19 February 2021, https://www.omidyargroup.com/wp-content/uploads/sites/7/2017/10/Social-Media-and-Democracy-October-5-2017.pdf

Tisne, M, 14 December 2018, 'It's Time for a Bill of Data Rights', *MIT Technology Review*, Accessed 19 February 2021, https://www.technologyreview.com/2018/12/14/138615/its-time-for-a-bill-of-data-rights/

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 Strategy and Plan of Action on Hate Speech, 2019, Accessed 16 February 2021, https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

UNODC, 31 March 2020, 'UN Tackles 'Infodemic' of Misinformation and Cybercrime in COVID-19 Crisis', *UNODC Department of Global Communications*, Accessed 23 February 2021, https://www.un.org/en/un-coronavirus-communications-team/un-tackling-%E2%80%98infodemic%E2%80%99-misinformation-and-cybercrime-covid-19

Waheed, A, 18 October 2015, 'Rape Used as a Weapon in Myanmar to Ignite Dear', *AlJazeera*, Accessed 16 February 2021, https://www.aljazeera.com/features/2015/10/28/rape-used-as-a-weapon-in-myanmar-to-ignite-fear

Wallstrom, M, 2015, 'Plugging Governments into Peace', *Building Peace*, Issue 5 '#Peacetech', Accessed 16 February 2021, https://creativeconomy.britishcouncil.org/media/uploads/files/Alliance_for_Peacebuilding_-_PeaceTech_Doc.pdf

Wood, M, 21 March 2019, 'Extremists Online: How a Troll Becomes a Terrorist', *Marketplace*, Accessed 19 February 2021, https://www.marketplace.org/2019/03/21/extremists-online-how-troll-becomes-terrorist/

Yanagizawa, D, 2012, 'Propaganda and Conflict: Theory and Evidence from the Rwandan Genocide', *Harvard University*, Accessed 15 February 2021, https://www.hks.harvard.edu/sites/default/files/centers/cid/files/publications/faculty-working-papers/257_Drott_Rwanda.pdf

Youngs, R, 11 September 2014, 'Digital Media Is a Double-Edged Sword', *Deutsche Welle, Global Media Forum*, Accessed 16 February 2021, https://carnegieeurope.eu/2014/09/11/digital-media-is-double-edged-sword-pub-56606