Cultural Violence and Peace Interventions in Social Media

18

Marc-André Kaufhold, Jasmin Haunschild and Christian Reuter

Abstract

Over the last decade, socio-technological innovations such as mobile technologies and social media services have strongly impacted modern culture and political processes. They are widely established in everyday life, but also relevant during natural and human-made crises and conflicts. For instance, Facebook was part of the 2010 so-called *Arab Spring*, in which the tool facilitated the communication and interaction between participants of political protests. Conversely, terrorists may recruit new members and disseminate ideologies. Based on the notions of cultural violence and cultural peace, this exploratory review firstly presents human cultural interventions in social media (e.g. dissemination of fake news, hate speech and terroristic propaganda) and respective countermeasures (e.g. algorithmic detection, counter-narratives, and reporting centres). Secondly, it discusses automatic cultural interventions realised via social bots (e.g. astroturfing, misdirection, and smoke screening) and countermeasures (e.g. crowdsourcing and visual analytics). Finally, this chapter proposes

M.-A. Kaufhold (✉) · J. Haunschild · C. Reuter
Science and Technology for Peace and Security (PEASEC),
Technische Universität Darmstadt, Darmstadt, Germany
e-mail: kaufhold@peasec.tu-darmstadt.de

J. Haunschild
e-mail: haunschild@peasec.de

C. Reuter
e-mail: reuter@peasec.tu-darmstadt.de

C. Reuter (ed.), *Information Technology for Peace and Security*,
Technology, Peace and Security I Technologie, Frieden und Sicherheit,
https://doi.org/10.1007/978-3-658-44810-3_18

to differentiate a range of cultural interventions in terms of actors (human vs. machine) and intentions (conflict vs. peace) to identify future research potentials for supporting situational assessments during conflicts.

Objectives

- Being able to describe and differentiate the complementary notions of direct, structural and cultural violence and peace, and to understand their relation to social media.
- Understanding definitions, classifications and use cases of social media, social bots and supportive ICT.
- Being able to distinguish how cultural interventions both by social media users and social bots may support conflicts but also promote societal peace.

18.1 Introduction

Mobile technologies and social media have enabled enormous socio-technological innovations with significant impacts on modern culture and political processes. **Social media** are used by citizens, journalists, organisations, political groups and businesses for a variety of purposes. This has led to a democratisation of public discourses, with actors gaining access to new audiences, being able to better target their information and to coordinate activities (Reuter & Kaufhold, 2018). Large-scale international conflicts or uprisings, such as the 2010 *thawra* (often referred to as *Arab Spring* (Avery, 2021)) showcased the potential of socio-technological transformations: Citizens were empowered by social media to coordinate protests and respond to crises (Reuter & Kaufhold, 2018). However, in other cases, the resulting reduction of state control and the spread of **false information** has also increased the complexity of tasks and put formal authorities under pressure. False information spreads quickly on social media and it is easy for groups to find an audience there, e.g. to enhance their profits or to target vulnerable groups with dangerous ideology. As such, social media is not only used for good or benign purposes[1]: Terrorists recruit new members and disseminate ideologies (Reuter et al., 2017), and social bots facilitate the dissemination of fake news and hate speech (Ferrara et al., 2016).

To understand the role of social media in promoting peace and conflict, the concepts of war, peace and security from the domains of Peace and Conflict Research and Security

[1] As the definition of good is a question of perspective, we do not claim universality. The opinion stated here and in the following is clearly our own moral conviction only.

Studies are helpful. They have identified the need to deepen and broaden the understanding of the relevant actors, objects of reference and threats (Booth, 2007). While in traditional research, the state was perceived as the central actor and the only object threatened, the conflict in the former Yugoslavia, for example, has shown that social groups can also be threatened by their own state and by other groups within the same state (Waever, 1993). This is particularly virulent in cyberspace, where it

> is also often unclear whether the actors pursue military-strategic or commercial objectives and whether they have no political, but maybe commercial interests maybe on behalf of the private sector or on behalf of a state or group with political intents. (Reuter, 2020, p. 13)

Similarly, the concept of **Human Security** shines a light on the potential threats to individuals, which do not only concern security aspects such as direct attacks, but also safety issues, such as health, development and environmental threats (Booth, 2007). This notion of the potential sources of harm and insecurity helps understand the role of social media as a socio-technological innovation, which, along with its emancipatory power, also amplifies existing threats. In this way, social media can contribute to direct, physical violence, e.g. through facilitating the recruitment of terrorists (Weimann, 2016), as well as to structural and cultural violence by creating, reinforcing and escalating grievances and political fragmentation, e.g. through the dissemination of fake news and of extremist ideologies (Reuter et al., 2017), partly aided by social bots (Stieglitz et al., 2017).

Accordingly, **socio-technological transformations** related to structural violence can be witnessed in a) the use and misuse of social media platforms to foster or erode intercultural understanding; and b) the use of **social bots** that can feign wide-spread support and amplify the spread of harmful content. However, innovations and regulations are also developed to mitigate socio-technological uncertainties in a way that curbs the misuse while maintaining the positive potential of social media. Based on the notions of cultural violence and cultural peace,[2] as proposed by Webel and Galtung (2007), this exploratory review presents human and automated cultural interventions in social media. Examples presented are the dissemination of fake news, hate speech and terrorist propaganda, as well as respective countermeasures, such as fake news detection, reporting centres and counter-narratives. Finally, this chapter discusses a range of cultural interventions in terms of actors (human vs. machine) and intentions (conflict vs. peace) to identify future research potentials for supporting situational assessments during conflicts.

[2] In peace and conflict research, there are different understandings of peace and violence. See Chapter 2 "*Peace Informatics: Bridging Peace and Conflict Studies with Computer Science*" for introductory explanations of the concepts around violence, war and peace.

18.2 Classifying Social Media Use

An interesting medium of the last decade are social networking sites, also called social media, which allow increased communication and collaboration among online users, and have become a ubiquitous part of everyday life for many citizens (Reuter & Kaufhold, 2018). **Social media** are often defined as a

> group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content. (Kaplan & Haenlein, 2010)

Research suggests that social media can be classified in terms of their "social presence/media richness" and "self-presentation/self-disclosure", allowing for diverse types of content exchange (see Table 18.1). Social media differ regarding the extent to which they are a virtual reflection of a person, with the reflection being enabled by higher media richness, e.g. in virtual social networks. In addition, these representations differ regarding the amount of self-presentation and self-disclosure, which is typically low in collaborative projects such as Wikipedia and high in virtual game worlds. These dimensions shape how virtual personas and digital relationships are perceived. The increasing presence of video and live streams leads to a higher perceived social presence and more trust.

Shaping opinions, politics, participation and protest, social media platforms are used by citizens for news consumption and social exchange (Robinson et al., 2017), by journalists for reporting, analysing and collecting information (Stieglitz, Mirbabaie, Ross, et al., 2018), and by organisations to monitor crises, emergencies, customer feedback and sentiment, amongst others (Haunschild et al., 2020). In this context, the research field of **crisis informatics** has emerged, which is a "multidisciplinary field combining computing and social science knowledge of disasters" (Soden & Palen, 2018, p. 2). However, due to some of social media's affordances, such as anonymity, depersonalisation and community cohesion, social media can contribute to cultural violence, for instance, spreading misinformation and disinformation commonly known as **fake news**, emphasising religious, ideological and linguistic divides as **hate speech**, or spreading propaganda in the case of **online terrorism**.

Table 18.1 Social media classification adapted from (Kaplan & Haenlein, 2010)

Social media		**Social presence/media richness**		
		Low	**Medium**	**High**
Self-presentation/ self-disclosure	**High**	Blogs	Social network sites (e.g. Facebook)	Virtual social worlds (e.g. Second Life)
	Low	Collaborative projects (e.g. Wikipedia)	Content communities (e.g. YouTube)	Virtual game worlds (e.g. World of Warcraft)

Table 18.2 Social bot classification adapted from (Stieglitz et al., 2017)

Social bots		Intent		
		Malicious	**Neutral**	**Benign**
Imitation of human behaviour	**High**	Astroturfing, conflict, doppelgänger, infiltration, influence, sybils	Humour	Chat bots
	Low	Spam, botnet command and control paying	Nonsense	News, recruitment, public dissemination, earthquake warning, editing and anti-vandalism

In social media, cultural interventions are not only disseminated manually by humans, but also automatically by social bots or large-scale botnets, which often act as multipliers (Yang et al., 2019). A **social bot** is

> a computer algorithm that automatically produces content and interacts with humans on social media, trying to emulate and possibly alter their behavior. (Ferrara et al., 2016, p. 96)

Bots' behaviour can establish realistic social networks and produce credible content with human-like patterns. Research suggests that social bots can be classified in terms of their malicious, neutral or benign intent, as well as a low or high level of human behaviour imitation (Stieglitz et al., 2017) (see Table 18.2). Even though these bots can be useful, for example in the context of improving citizen-generated information in case of crises and natural disasters (Maniou & Veglis, 2020; Stieglitz et al., 2022), they can also infiltrate political discussions, manipulate the stock market, steal personal information, or spread fake news. Thus, the use of bots facilitates the targeted spread of particular ideological content and views on social media, disguised as organic, natural human support, creating new socio-technological phenomena.

18.3 Case I: The Dissemination of Fabricated, Manipulated and Misinterpreted Content

Fake news has a long history, but due to its' increased spread and amplification in digital *echo chambers* and a resulting effect on societal opinion formation and politics (Becker, 2016), the term gained much more attention in the past years (Gregory, 2022; Reuter et al., 2019). However, fake news is difficult to categorize and the boundaries to interpretation of information are sometimes difficult to draw, inciting debate about the gatekeepers of true information and its online presentation. Currently, no agreed definition or conceptualisation of fake news exists, but many authors differentiate according to intent and content (Aïmeur et al., 2023).

18.3.1 Dissemination of Fake News in Social Media

While the term of fake news was originally used to refer to comedy news shows, in 2016 the perception changed when many fake stories went viral and started to affect political parties globally and impacted opinions on a larger scale than before (Becker, 2016). Yet, presenting news in a way that seeks to support a particular view is not a new phenomenon. Framing, the "persistent selection, emphasis, and exclusion" (Goffman, 1974, p. 7) of information is a common mechanism in news presentation, leading to the interpretation of information in a particular light. For example, migration has, in modern times, often been framed as a crisis rather than an opportunity (Georgiou & Zaborowski, 2017). In contrast to framing, which seeks to persuade by highlighting selected arguments, disinformation intentionally deceives (Volkova & Jang, 2018). A further difference consists in the degree to which information is falsified or presented in a misleading way (see Fig. 18.1).

Allcott and Gentzkow (2017, p. 213) define fake news as "news articles that are intentionally and verifiably false and could mislead readers" and distinguish it from similar phenomena like unintentional reporting mistakes, rumours, conspiracy theories, obvious satire, and more. Similarly, Sängerlaub (2017a) defines fake news as intended disinformation and describes three types of fake news. First, there is completely fictitious news which he refers to as **fabricated content**. For example, segments from video games have been used to, purportedly, show scenes of war and fighting (Tagesschau, 2023, see Fig. 18.2).

Second, **manipulated content** is based on accurate information, which is manipulated in some respects. Instead of inventing new content or media material, existing material is used and displayed in a manipulative manner. The use of artificial intelligence has enabled the creation of fabricated content based on pictures and voice segments that are available online. However, rather than creating completely new content, existing material is usually used to increase believability and quality. For example, in the Russian invasion of Ukraine in February, 2022, a picture was altered to purportedly show drugs on the Ukrainian President's desk (Euronews, 2022). In addition, a video was altered to, falsely, show the Ukrainian President asking citizens to surrender (ibid.).

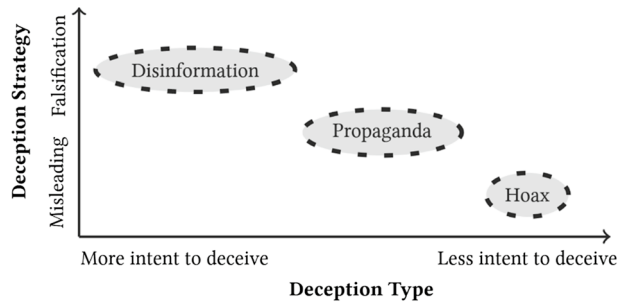

Fig. 18.1 News categorised based on deception type and strategy. Source: Volkova & Jang, 2018, p. 576

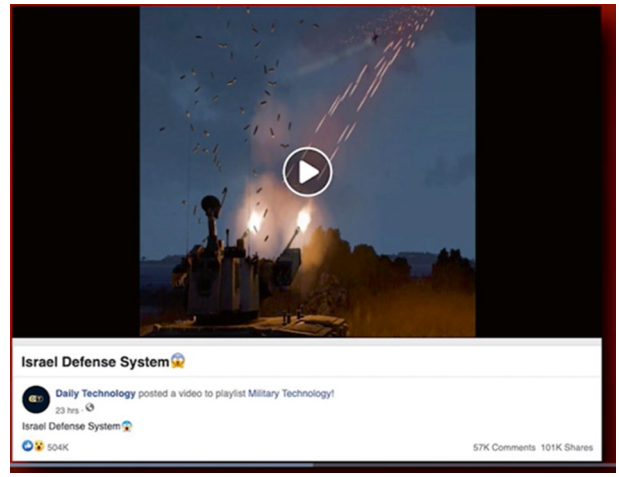

Fig. 18.2 Segments of video games (here "Arma3") used to purportedly show scenes of fighting in current conflicts, such as in Ukraine or Gaza. Source: Stern, 2023

Third, **misinterpreted content** refers to correct information which is quoted out of context or is intentionally misinterpreted by the author. For example, a video of Uzbek soldiers dancing at a military concert in Tashkent was used as pro-Russian propaganda by integrating a header which claimed that the video showed Russian soldiers joyful at the prospect of going to war, even though the video could be found on the web long before the invasion (see Fig. 18.3). Similarly, older pictures and videos from other conflicts or accidents, or from military drills are often used and claimed to show current incidents (Deutsche Welle, 2022). Another strategy involves claiming that opposing conflict parties are staging attacks and atrocities. This was the case in Syria, for example, where Russian media falsely reported that the gas attack on Duma in 2018 had been staged. This claim was made by showing pictures of the shooting of the film "*Revolutionary Man*" prior to the attack (Fig. 18.3, Tagesschau, 2018).

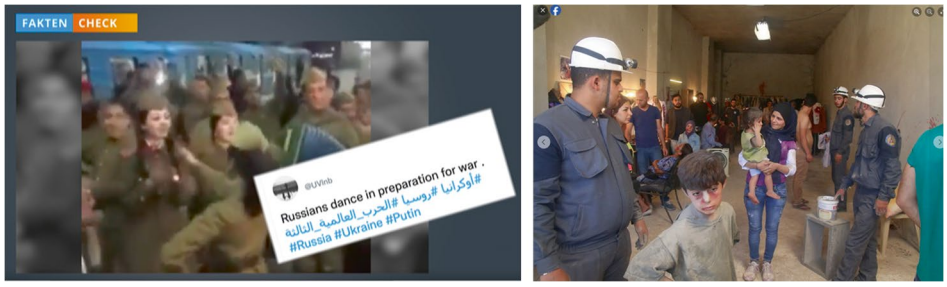

Fig. 18.3 Left: Video claiming to show Russian soldiers before deployment to Ukraine, but, in reality, showing Uzbek soldiers dancing at a concert. Source: Deutsche Welle, 2022. Right: A picture from a film set used to claim that White Helmet volunteers in Syria were staging atrocities. Source: Snopes, 2018

The topics covered by fake news are often negative and controversial, such as migration, child abuse, or war, arousing high emotions (Ziegele et al., 2014). However, prevalent types of fake news differ between states and cultures (Humprecht, 2019). Furthermore, fake news can have serious consequences, e.g. influencing elections, stock markets or leading to direct violence (Kaufhold & Reuter, 2019). In an illustrative case in South Africa in 2019, foreign shops were attacked, leading to the deaths of 12 people, mostly nationals, while tensions between South Africans and Nigerians increased with footage on social media from different times and places falsely claiming to portray attacks against Nigerians (Chenzi, 2021). This case shows how already existing xenophobia and grievances can be exacerbated by social media, leading to retribution for violence that did not actually occur. In 2018, orchestrated by the Myanmar military, intense violence amounting to ethnic cleansing erupted against the Muslim minority of Rohingyas living in Myanmar. Investigations have shown that social media accounts with a following of 1.3 million ostensibly dedicated to entertainment had been set up by the military and used to sow hatred against the minority.

> The […] actions by Myanmar's military on Facebook are among the first examples of an authoritarian government's using the social network against its own people. […] Troll accounts run by the military helped spread the content, shut down critics and fuel arguments between commenters to rile people up. Often, they posted sham photos of corpses that they said were evidence of Rohingya-perpetrated massacres. (Mozur, 2018)

The strategy involved spreading rumours to both sides, Muslim and Buddhist, about imminent violent attacks by the other group with the aim of spreading insecurity that would increase the populations' reliance on the military. In addition, the strategy also included discrediting users who posted content critical of the military.

As shown by the examples above, fake news has also been an element in the invasion of Russia in Ukraine in 2022. While spread by both sides, Russia has control over the media and has been leading a state-imposed propaganda campaign, including elements of fake news (Khaldarova & Pantti, 2020). The strategy has encompassed associating Ukrainians with fascists and portraying the West as an aggressor (Khaldarova & Pantti, 2020; Rossoliński-Liebe & Willems, 2022). Another example of the power of false information has been the US presidential election in 2020. In 2021, the storming of the United States Capitol occurred after then-president Donald Trump delivered a speech in which he repeatedly claimed that the 2020 election, won by his competitor, Joe Biden, was fraudulent and encouraged his followers to "fight". This event was preceded by over 1,500 tweets from Trump containing the aforementioned claim in the months leading up to it (Fuchs, 2021).

In addition to financial motives (Klein & Wueller, 2017), ideological motivations are relevant (Allcott & Gentzkow, 2017), with fake news used to manipulate public opinion and debate. Well-known incidents are the US presidential election in 2016 (McCarthy, 2017) and the UK Brexit referendum, where false information were often employed in combination with social bots (Mostrous et al., 2017). In times of the heightened num-

bers of refugees and the prevalence of right-wing populism, fake news in Europe often deals with migration and refugees. According to research by the German investigative journalism collective Corrective, most fake news in Germany originated from supporters and politicians of the right-wing populist party Alternative für Deutschland (AfD). The party's attitude becomes explicit in the statement of its spokesman Christian Lüth:

> If the message fits, we actually don't care where it comes from and how it was created. It's no big deal if it's fake. (Faktenfinder, 2017)

In 2018, Facebook removed numerous accounts and pages for spreading hate speech and false news about the Rohingya community in Myanmar. Among the deleted accounts was that of Min Aung Hlaing, who served as Commander-in-Chief at the time and, after the coup, assumed the role of Prime Minister in 2021. However, the Burmese government, which has faced accusations of genocide against the Rohingya people, denied any involvement in the incident (Kyaw, 2019). Another instance of political manipulation employing fake news can be seen during the Russian aggression against Ukraine in 2022. Russian President Vladimir Putin falsely accuses Ukrainians of perpetrating genocide against Russian-speaking communities in eastern Ukraine, creating a fabricated scenario of threat. He then refers to his attack as a "denazification" of Ukraine (Rossoliński-Liebe & Willems, 2022).

Furthermore, compromised accounts, which have been taken over by attackers temporarily or entirely through **account hijacking**, are sometimes used to disseminate fake news. Usually, human attackers or programmed bots obtain users' login details via phishing, malware, or cross-site scripting. Existing as viruses, malware can replicate itself by sending links or direct downloads to other social media users. Account hijacking can be used for political purposes, with compromised accounts being, due to their relationships of trust with legitimate users, more valuable than bots regarding the distribution of misinformation and propaganda (Trang et al., 2015). In X (formerly known as Twitter), for instance, social bots can act as **fake followers** or disseminate **fake retweets**, which are motivated by the fact that a high number of followers and retweets suggest popularity and high reputation (Jiang et al., 2016; Wu et al., 2015). There are examples of politicians and celebrities buying fake followers to gain more popularity statistically and increase their value on X (Jiang et al., 2016). Using fake retweets, it is possible to create popularity and broaden the audience artificially (Wu et al., 2015). Fake retweets and followers are often purchased on online marketplaces; fraud is conducted with the help of bots or malware-infected accounts.

Another threat to society is posed by hyper-realistic videos produced through Generative Artificial Intelligence, commonly known as deepfakes. These manipulated videos allow people to create false representations of events that never took place (Westerlund, 2019), for example by replacing the face of a speaker by that of another person, or by synthesising the voice of another person (Godulla et al., 2021). The combination of such videos with previously discussed dissemination practices can result in highly convincing fake news and further erode the credibility of legitimate news content.

18.3.2 Countermeasures Against Fake News

So far, there is no clear answer to what the most appropriate approach on how to tackle fake news looks like. Identifying solutions and responsibilities to prevent individuals and society from possible negative effects is a complex task. Nonetheless, researchers have presented several approaches to detect and handle fake news. Three enablers and corresponding response vectors have been identified for countering fake news: To address the susceptibility of the host (news readers and social media users), education and clarification is the most promising avenue. Another enabler is a conducive environment, consisting of toxic and complicit platforms, which can be addressed through regulation. Finally, the various types of fakes acting as virulent pathogens can be addressed through auto-detection (Rubin, 2019). Focusing on different strategies, Verstraete et al. (2022) describe laws, markets, code-based interventions and norms as possible angles for limiting fake news.

Reviewing the literature, we deduce five possible approaches to **countering fake news** (Table 18.3). Most social networks have taken measures such as curating, deleting and censoring. In doing so, even initially independent platforms now take the traditional journalistic role of **information gatekeeper** (Wohn et al., 2017). Many platforms provide mechanisms for users to flag content that they believe to be false (Ng et al., 2021). These annotations are then checked by experts, belonging either to the platform or to national independent fact-checking organisations. This expert-oriented checking of facts

Table 18.3 Measures against fake news in social media. (Source: Own depiction)

Gatekeeping	Gatekeeping is the process through which information, including fake news, is filtered for dissemination, e.g. for publication, broadcasting, social media, or some other mode of communication (Barzilai-Nahon, 2009)
Crowd-Sourced Content Moderation	Through crowd-sourced assessments, the "wisdom of the crowd" can be used to evaluate the veracity of content, correct it or provide it with context (Wirtschafter & Majumder, 2023; Wojcik et al., 2022)
Media Literacy	The purpose of media literacy – a multi-dimensional process allowing people to access, evaluate and create media content – is to help people to protect themselves from the potentially negative effects of (mass) media (Potter, 2010)
Law/Regulation	Laws may assist in fighting fake news and hate speech by sanctioning platforms that disseminate fake news or hoaxes by penalising them or by forcing them to quickly delete illegal contents; however, laws potentially threaten freedom of speech (Miró-Llinares & Aguerri, 2023; Müller & Denner, 2017)
Algorithmic Detection	The algorithmic detection of fake news comprises classification-based (e.g. machine learning), propagation-based (e.g. social network analysis) and survey-based approaches (Viviani & Pasi, 2017)

is based on human work and deals with the exposure of false statements. The experts check their researched and already created lists with the articles flagged by Facebook users.

As another approach for verification, **crowd-sourced content moderation** is employed on several social networking sites, such as Wikipedia, YouTube, Reddit and X (Wirtschafter & Majumder, 2023). Empirical data shows that flagging fake news after they are detected reduces the reach of fake news inside the network (Ng et al., 2021). Instead of experts such as journalists, social media users assess and comment on the veracity of posts. It often involves a prioritisation of trusted moderators who have a history of positive and particularly helpful contributions (Wirtschafter & Majumder, 2023). Since 2022, Community Notes can be added to posts on X to correct it or provide context (see Fig. 18.4). These notes can be judged by others as helpful or unhelpful, and this statement is locked and thus be permanently attached to a note when it receives enough congruent judgements from people who have previously disagreed about other notes (Wirtschafter & Majumder, 2023). However, a study indicates that political partisanship significantly influences which posts users challenge or which notes they rate as unhelpful (Allen et al., 2022). In addition, previous work has found both machine learning algorithms as well as crowdsourcing to be less accurate than professional fact checking and to work better with politically educated people (Godel et al., 2021).

In addition, technological means are used to limit the visibility of fake news on social media by reducing their relevance in news feeds and to limit their spread, e.g. reducing the amount of possible forwarding on messenger apps to five (Hern, 2020, Ng et al., 2021). The Chinese social network Sina Weibo relies on social reporting of fake news and penalizes users' posting and sharing of false information by reducing users' points (Ng et al., 2021). When users' points fall below a threshold, all their posts are automatically blocked from being able to be shared (Ng et al., 2021).

Fig. 18.4 Left: Example of Community Notes on Twitter during pilot testing. Source: Wojcik et al., 2022. Right: Fake News Assessment Page from Sina Weibo. Source: Ng et al., 2021

Furthermore, efforts are made to increase the populations' **media literacy**. Research suggests that people with good media literacy are better able to navigate through today's media age and to identify and critique false news but also to create fake news themselves (Mihailidis & Viotty, 2017). The ability to proficiently use media for one's own goals and needs is an integral part of removing the influence of fake news and general misinformation as well as preventing its spread (Cooke, 2017). One aspect that helps people recognise false information is the style of the information (Hancock et al., 2008). Since fraudsters do not present accurate information but invent it, they have to be creative and use their inventive abilities. Hancock et al. (2008) found that fraudsters rely on more sense-based words, less self-oriented and more other-oriented words. In addition, positive emotions in a text lower the probability of news being fake (Nanath et al., 2022). A study has found that different types of false information trigger different emotions, e.g. propaganda triggers extreme positive and negative emotions, whereas Satire invokes disgust and clickbait surprise (Ghanem et al., 2020). *Neue Wege des Lernens e.V.* (2017), a registered association in Germany, developed an app called Fake News Check. The app is designed to help users ask the right questions and distinguish fake news through guided reflection from real news. By asking 19 questions about a news item, the app aims to sensitise for the critical handling of news.

At the beginning of 2018, the European Commission appointed a High Level Group on fake news and online disinformation consisting of 39 experts from science, media, and social media platforms. Just before, in October 2017, a German law came to force called *Netzwerkdurchsetzungsgesetz* (NetzDG, Network Enforcement Act). It attempts to fight fake news and hate speech by forcing platforms to delete illegal contents quickly. However, it has been widely criticised for threatening freedom of speech, although there are also voices endorsing the law for supporting the victims of fake news and hate speech. Müller and Denner (2017) state that deleting fake news from social networks is not the best solution. Instead, it would create reactance, an even more fertile ground for conspiracy ideas and the tendency to social divide. They argue that the NetzDG threatens freedom of speech by forcing social networks to delete content pre-emptively, if there is any suspicion of fake news. Furthermore, laws could also be established to prevent advertising revenues for clickbait websites that use fake news and hoaxes (Klein & Wueller, 2017).

There are several approaches to algorithms and systems which facilitate **fake news detection.** Assistance tools, such as TrustyTweet (Hartwig & Reuter, 2019), TweetCred (Gupta et al., 2014) or Bot-Detective (Kouvela et al., 2020) help users identify fake news and bot-driven accounts. Similarly, Narwal et al. (2017) presented an assistant system supporting the detection of visual bias in images. It facilitates users in detecting biases and sharing their findings on Twitter. Furthermore, the system comprises bots engaging affected users into a conversation about the bias. In a comprehensive review, Viviani and Pasi (2017) compare different algorithms for fake news detection, distinguishing classification-based (including machine learning), propagation-based (including social network analysis) and survey-based (including representative samples) approaches.

These approaches place the responsibility for dealing with disinformation on different groups. Media literacy targets the recipients of fake news. These can be aided by the

inclusion of additional information that supports them in identifying fake news, such as adding crowd-sourced flags or information about the political alignment of their news feed (Behzad et al., 2023). In contrast, regulation demands that either governments or social media platforms make and enforce rules about limiting the availability or spread of fabricated content. Gatekeeping can be performed either by experts employed by social media platforms or by journalists organised in independent fact-checking institutions (Graves, 2018). Their results can either prevent fake news from being shown, can be used to inform consumers or to reduce the sharing and visibility of posts that are suspected of spreading false information. Similarly, algorithmic solutions support any of the actors, pointing out identified fake news either to media consumers, to platforms, gatekeepers or regulators, depending on who is deemed responsible. While citizens are undecided about who should take that responsibility, the majority of Germans support relevant authorities' swift reaction to fake news, but also transparent journalism (Reuter et al., 2019).

18.4 Case II: Cyber Abuse as a Vehicle of Violence Against Individuals and Groups

Besides fake news, citizens and professionals are increasingly exposed to **digital violence**, such as cyberbullying and **hate speech** (Kaufhold et al., 2023). In German debates, the meaning of fake news and hate speech is often mixed, although they represent different phenomena (Sängerlaub, 2017b). While the internet has now produced a variety of cyber abuse awareness, reporting and prevention campaigns for end-users, law enforcement agencies are deployed in many countries and organisations to enhance the preventive and reactive capabilities against cyber abuse. Still, the amount of cyber abuse context is increasing, and the tasks of law enforcement agencies are becoming more complex due to the increasing amount and varying quality of information disseminated into public channels.

18.4.1 Cyber Bullying and Hate Speech in Social Media

Cyber abuse phenomena increasingly arise from digital space, including cyber bullying and hate speech. **Cyber bullying** means "insulting, threatening, exposing or harassing people using communication media, such as smartphones, emails, websites, forums, chats and communities" (BMFSFJ, 2022). While cyberbullying is mostly directed against individuals, hate speech usually refers to groups of people. According to the European Commission against Racism and Intolerance, hate speech includes

> all forms of expression that denigrate, belittle, insult, stigmatise, threaten or attack people or groups of people on the basis of perceived group-related characteristics and status characteristics attributed to them. (ECRI, 2015)

Against the background of an increasingly complex information space, special framework conditions arise with regard to civil security.

According to a comparative study by the Bündnis gegen Cybermobbing e.V. (Beitzinger & Leest, 2021), around 12% of the German population were affected by cyberbullying in 2021. While slightly more than 53% of cyberbullying incidents occur in the private sphere, 38% still occur in a work environment. In addition to depression, addiction risk or physical complaints, around 15% of those affected by bullying and cyberbullying classified themselves as suicidal. While over a third of those affected had communicated with friends or family in response to (cyber)bullying, another third said they had taken no action and only 15% said they had looked for information and help on the internet. From an economic point of view, the willingness of bullying victims to quit is 40% higher, those affected have almost twice as many sick days as the average and the annual costs of lost production in the German economy are estimated at around eight billion euros.

Hate speech is also pervasive and it mainly targets disadvantaged or minority groups. Banaji and Bhat (2021, p. 21) suggest that hate speech has particularly racist, sexist and misogynist, xeno-, homo- and transphobic content, classist or caste-based, and ageist content. Similarly, a systematic literature review establishes the categories of online religious hate speech, identifying particularly Islamophobic hate (Castaño-Pulgarín et al., 2021), often triggered by acts of terrorism. However, antisemitic online hate is also pervasive, partly related to the Israel-Palestinian conflict, but also mingles with racist and anti-capitalist stereotyping and conspiracy theories (Bundeszentrale für politische Bildung, 2020). Other types are online racism against Indigenous peoples and People of Colour, political online hate, which tends to intersect with fake news and conspiracy theories, and gendered online hate (Castaño-Pulgarín et al., 2021) (see Fig. 18.5).

A regular survey by the Media Authority of North Rhine-Westphalia (Landesanstalt für Medien NRW, 2021) shows that the number of internet users in Germany who are frequently confronted with hate speech has risen in recent years from 27% (2017) to 39% (2021). Although more than two-thirds of the respondents in 2021 have already noticed hate comments, only 28% of them have reported a hate comment to the respective portal. Nevertheless, internet users see prosecution (87%) or deletion of hate comments (73%) as more effective than behavioural guidelines (42%) or active **counter-speech** (17%).

The dissemination of hate speech is sometimes supported by paid authors, fake accounts or social bots, for instance, as **astroturfing** campaigns, which describes pretending to constitute a grassroot[3] movement to use the image of a local, social initiative or organisation to influence economic or political conditions (Cho et al., 2011). It aims

[3] Grassroot organisations are defined as "local political organizations which seek to influence conditions not related to the working situation of the participants and which have the activity of the participants as their primary resource." (Gundelach, 1979, p. 187).

Fig. 18.5 Left: Intersection of antisemitic online hate and conspiracy theory based on a meme depicting a heavily stereotyped Jewish man, used by the alt-right, in circulation online since approx. 2004. Source: Oboler, 2014. Right: Nazi image of Winston Churchill

at manipulating people's (political) opinions by strengthening their own views or discrediting contrary arguments by expressing doubts or neglecting arguments. An analysis indicates that over 100.000 fake and compromised accounts are used for astroturfing on Twitter, accounting for 20% of the top ten global trends (Elmas et al., 2021). Instead of targeting the outcome of a particular policy, the Russian bot firm Internet Research Agency (IRA) was used to manipulate voters in the 2016 US election (Diresta et al., 2019). It had set up accounts across all main social media platforms and used astroturfing to, among other things, encourage and discourage certain voter groups. Research shows that the bot firm co-opted debates such as the #BlackLivesMatter movement and spread posts on the extreme spectrum of both right and left positions, using existing grievances to increase fragmentation, societal insecurity and distrust in the democratic institutions (Stewart et al., 2018). However, non-state groups active on social media are a very heterogeneous group and their possible financiers and motives can be difficult to establish, making it hard to determine their legitimacy and claim for representing. Case studies show that some groups' bot-like activities can amount to political manipulation – such as right-wing online politics of part of the Hindu nationalist diaspora (Mohan, 2015) or the Iranian diasporic group Mojahedin-e Khalq (MEK). MEK seeks to influence US and EU foreign policy related to Iran by mobilising "international human rights of Middle Eastern women […] toward Western militarist agendas" (Honari & Alinejad, 2022, p. 919), amounting to "a tactical performance of civic participation" (Honari & Alinejad, 2022, p. 920).

18.4.2 Strategies and Technologies for Dealing with Cyber Abuse

When it comes to cyber abuse, in part strategies are similar to those of fake news detection (see Table 18.4). For example, education and deletion also play a role when it comes

Table 18.4 Measures against cyber abuse (in addition to measures similar to countering fake news). (Source: Own depiction)

Networking Centre	Networking centres connect actors with initiatives that provide relevant services and support, such as help for victims, education and awareness, skill development (Iginio et al., 2015)
Reporting Centres	Reporting centres facilitate the reporting of hate speech, provide counselling and support services for affected citizens, forward comments to responsible authorities (such as law enforcement agencies), and send delete requests to platforms (Kaufhold et al., 2023)
Visual Analytics	Visual analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of large and complex data sets (Keim et al., 2008), which can be used by reporting centres

to hate speech and cyber bullying (Citron & Norton, 2011). **Educational measures** can help raise citizens' awareness, offer support in developing creative solutions against hate speech (Iginio et al., 2015). The information portal DAS NETTZ, for example, is a networking centre against hate speech and offers a search for initiatives from German-speaking countries that can be filtered by topics such as de-escalation, counter-speech, support or reporting hate speech (Das NETTZ gGmbH, 2023).

Similar to fake news, the removal of hate speech in Germany is primarily defined by the NetzDG, which requires social network operators to remove or block "obviously illegal content within 24 h" of receiving a complaint (§ 3 Abs. 2 Nr. 2 NetzDG). As part of the HessenGegenHetze (Hesse against hate) initiative, the state has established a **reporting centre** for citizens (HMdIS, 2022). This office serves to provide counselling and support services to those affected by hate comments, while also forwarding these comments to platform operators with the aim of quickly removing hate speech from public perception (Kaufhold et al., 2023). The voluntary initiative Hassmelden (Reporting Hate) (discontinued in 2022 due to the heavy overburdening with cases) was one of the first and only central reporting office for hate speech, which also offered a smartphone app for reporting hate speech (Hassmelden, 2022).

Due to the significant psychological and reputational costs of cyber bullying and the disruptive effects of hate speech, these instances can be persecuted by the police. In contrast to false and misleading information, due to the history of Holocaust revisionism, some aspects of hate speech are relatively clearly defined in Germany and can be similarly applied and prosecuted in the digital domain as it is to the analogue world. Therefore, reporting needs to pay attention to the judicial requirements for using social media posts as evidence in trials (Kaufhold et al., 2023). However, the governance of hate speech differs between countries, with some focusing more on penalisation and others on social media platforms' corporate social responsibility (Doncel-Martín et al., 2023).

Hate speech and supporting fake accounts (Schoch et al., 2022) can also been identified through **algorithmic detection**. In principle, many algorithms have already been tested and datasets published that enable automatic detection of cyberbullying (e.g. Elsafoury et al., 2021) and hate speech (Fortuna & Nunes, 2018; Poletto et al., 2021) in social media using AI, especially artificial neural networks. Current research suggests that classification quality can be improved by using large language models (Chiu & Alexander, 2021). Flexibility can also be improved by adapting those models with Few-Shot Learning, i.e. using a small domain-specific training data set. As quantity and quality of data become increasingly important to further improve the classification quality of models (Bayer et al., 2022; Rizos et al., 2019), the research area of data augmentation investigates the artificial generation of training data (Feng et al., 2021).

However, uncritical data annotation and model building can lead to cyberbullying (Gencoglu, 2021) and hate speech (Mou & Lee, 2021; Sap et al., 2020) detection algorithms reinforcing social biases (Solaiman et al., 2019). Furthermore, automatic hate speech detection faces the problem of overfitting and thereby a lack of generalisability due to aforementioned biases and because hate speech changes with time (Yin & Zubiaga, 2021). Thus, existing research has examined enhanced practices of **crowdsourcing** for an improved labelling of abusive behaviour (Founta et al., 2018). Furthermore, research shows that interpolation-based approaches can mitigate this effect (Chen et al., 2020; Shi et al., 2021). For this, it is essential that users can understand the decisions made by the algorithm. The use of model-agnostic white-box approaches, such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) seems promising to explain and visualise these decisions.

After the classification of the data, an appealing and target-oriented visualisation of the situation is still required in order to establish appropriate situational awareness and to support the decision-making based on it (Eismann et al., 2018; Zade et al., 2018). The sheer amount of data, also called Big Social Data (Olshannikova et al., 2017), that is generated in everyday life and during major events across platforms, for example on Facebook, Telegram or X, can lead to information overload, which implies that technical support solutions must have very good usability as well as configurable filter mechanisms and classifiers in order to reduce the amount of data (Kaufhold, Rupp, et al., 2020). To facilitate the analysis, **visual analytics** approaches combine

> automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets. (Keim et al., 2008)

While crisis informatics has explored interactive interfaces for the collection and analysis of public data for crisis management (Kaufhold, Bayer, et al., 2020; Onorati et al., 2019), there are only a few research approaches for the visualisation of cyberbullying (López-Martínez et al., 2019) and hate speech (Bunde, 2021; Paschalides et al., 2020), which are not tailored to the requirements and needs of law enforcement agencies.

18.5 Case III: Propaganda and Recruitment in the Realm of Online Terrorism

As indicated, the spread of disinformation is strongly driven by the motivations of different actors. The recent past saw an increase of terrorist attacks across Europe, such as the November 2015 Paris attacks, the 2016 Brussels bombings or 2017 London bridge attack (Stieglitz et al., 2018). Besides **direct violence**, the internet and especially social media are also used to promote **cultural violence**, e.g. by disseminating ideologies of terrorism and recruiting new members. Again, radicalisation and recruitment into terrorist and extremist organisations is only possible where terrorist propaganda meets experiences or perceptions of injustice and grievances (Al-Saggaf, 2016). Research indicates that the majority of terrorist are recruited offline, and that offline recruits are more likely to attack and their attacks are more deadly (Hamid & Ariza, 2022). In addition, research stresses the interconnectedness of the online and offline realm, for example with radical online content being consumed together in the community, or with radical content from the community being shared and discussed online (Whittaker, 2022).

18.5.1 Propaganda and Recruitment in Social Media

As for research on **terrorist organisations** and social media in general, much of the research in this field deals with the so-called Islamic State (IS, ISIS, ISIL, DAESH). Media plays a significant role in terrorism since terrorism can only gain importance if it becomes meaningful on the media level:

> Without a letter of confession, a farewell video by the assassin or a last posting in the social network a bomb attack would be nothing else than a capital crime. Only through the terrorist communications strategy, the crime turns into a terrorist act. (Christoph, 2015, p. 145)

However, terrorists do not rely on media-makers, but have themselves become agents in social media. Social media offer the advantage of immersion, which means the merger of medium and message, and the credibility of terrorist narrations is strengthened by spreading it on established platforms like YouTube (Christoph, 2015).

Klausen et al. (2012) stress that the British terrorist group al-Muhajiroun uses its international network of YouTube-channels elaborately for propaganda and the presentation of violent content. Social media are used to incite phantasies and to normalise extreme views by creating an echo chamber of like-minded individuals (Awan, 2017; Torok, 2015). Weimann and Jost (2015) explain the use of Facebook, X, and YouTube by terrorist organisations for recruitment and propaganda: social media make it easier to find like-minded people and to consume their online content as it

> provides a stage on which ISIS can perform its recruitment-oriented 'theater', presenting a carefully packaged image of itself as the fulfilment of a kind of ultimate jihadi fantasy. (Torok, 2015)

Thus, social media constitutes an institution wherein extreme beliefs and actions are normalised, or made to seem the standard practices of dedicated Muslims (Torok, 2015). This leads to ISIS developing and disseminating its central narratives, often by reframing familiar concepts such as jihad and martyrdom (Torok, 2015). By performing this jihadi fantasy of normalised extremism, ISIS encourages young Muslims to follow them as a family.

Simultaneously, terrorists can address an almost endless number of potential members via social media, who would otherwise not find the way to closed forums, which were primary points of contact for members, interested parties, and newcomers in the past (Weimann & Jost, 2015). Weimann adds that other online services are also involved in the **recruitment** and radicalisation process, "such as Kik or Skype [which] allow for direct, real-time communication between recruiters and their audiences" (Weimann, 2016, p. 82). Another aspect is the professionality in handling social media. The members' language and translation skills contribute to the facilitation of understanding (Gates & Podder, 2015). Also, the IS propaganda performed well with respect to recruiting not only potential new fighters, but also technically proficient and talented users of social media to sustain recruitment (Gates & Podder, 2015). Since May 2014, IS videos or other media have been produced by the al-Hayat Media Center, a special production unit for Western recruitment (Weimann, 2016). The materials by al-Hayat Media Center exist in many languages and are spread via social media. For example, "IS released a video inciting Muslims to come and participate in jihad, featuring a German chant with an English translation" (Weimann, 2016, p. 80). In the Israeli-Palestinian conflict, research indicates that the propagation of mobilising content across Palestinian social network sites played a significant role in the occurrence of several lone-wolf terrorist assaults that targeted Israeli civilians between October 2015 and September 2016 (Chorev, 2019).

Often in combination with social bots, **social spam** is utilised for political purposes, aiming at the distribution of wrong and confusing information as well as prevention and complication of communication among users, e.g. conversations about recent political events (Almaatouq et al., 2016). Thus, spam is often used to manipulate social media users' perceptions of relevant issues. Performing misdirection, posts referring to a certain hashtag are spammed for distraction. Then, users perceive posts making other issues subject to discussion, shifting focus away from genuine topics of public interest. For example, a Syrian botnet distributed tweets to diverse events, independent of the hashtag used as a reference point (Abokhodair et al., 2015). In contrast, **smoke screening** entails the process of tweeting referring to a certain topic or hashtag to make identifying potentially relevant posts more difficult for the perceiving users. Syrian bots also applied this tactic to overwhelm pro-revolutionist tweets under the hashtag "#Syria".

18.5.2 Counterterrorism in Social Media

A variety of different measures to counter terrorism have been identified in research (see Table 18.5). Gartenstein-Ross (2015) opens up a new perspective on terrorist actions on

Table 18.5 Measures against terrorism in social media. (Source: Own depiction)

Clarification	Clarification means trying to answer to terrorist propaganda with logic to invalidate it, i.e. statements, which clarify unknown connections (Reuter et al., 2017).
Counter-Narratives	A narrative that goes against another narrative. Narratives are compelling storylines which can explain events convincingly and from which inferences can be drawn (Freedman, 2006).
Parody/Satire	Parody is a hilarious satirical imitation by distortion and exaggeration. Satire is a genre which criticises and stultifies events. Both aim at expressing mockery about serious issues (Reuter et al., 2017).
Hacking	Hacking refers to legal and illegal activities, such as the blocking of accounts and the appeal to the population to report suspected persons as well as activities by multiplying parodist media (Reuter et al., 2017).

the internet: He concedes that IS uses social networking sites such as Twitter successfully, but simultaneously draws attention to the fact that IS also relies on the success of this propaganda and is thus vulnerable to disruptions of this communication. A further study contributes explorative insights on the fight against terrorism in social media, especially on Twitter (renamed to X) (Reuter et al., 2017). By applying qualitative content analysis on anti-propaganda in tweets and by comparing terrorists' statements to expressions of the US government or media reports, they identified three categories of countermeasures: **clarification**, **parody/satire**, and **hacking**. The study concludes with the recommendations to start mass movements, convey authenticity and credibility, use parody and satire for critical reflection, promote resistance on eye level, perform hacking by specialised groups, and to convey understandable clarification. Satirical content is shown to receive most attention, while the success of hacking scenes is judged as limited due to the ease of reopening accounts and moving content to other platforms.

Jeberson and Sharma (2015) focus on determining possible methods to identify terror suspects in social networks. Cheong and Lee (2011) suggest the establishment of a knowledge base in connection with intelligent data mining, visualisation and filter methods, allowing authorities and decision-makers a quick reaction and control during terrorist scenarios. Furthermore, Weinmann and Jost (2015) suggest that the analysis of terrorist online communication can provide insights into the way of thinking, the motivation, the plans, and fears of terrorist groups. Instead of strict censorship of radical contents, terrorist communication strategies should be disturbed by a mixture of technical (e.g. hacking) and especially psychological (e.g. anti-propaganda) means (Weimann & Jost, 2015). Hussain and Saltman (2014) emphasise that general censorship can be counterproductive and suggest positive measures such as expanding contents against extremism. Other initiatives focus on prevention through (offline) information at schools, universities or prisons (Saltman & Russell, 2014). Weimann (2016) sees the governments, researchers, and the wider security community in the role of a counterterrorism force. For the security community, according to Weimann, it is necessary

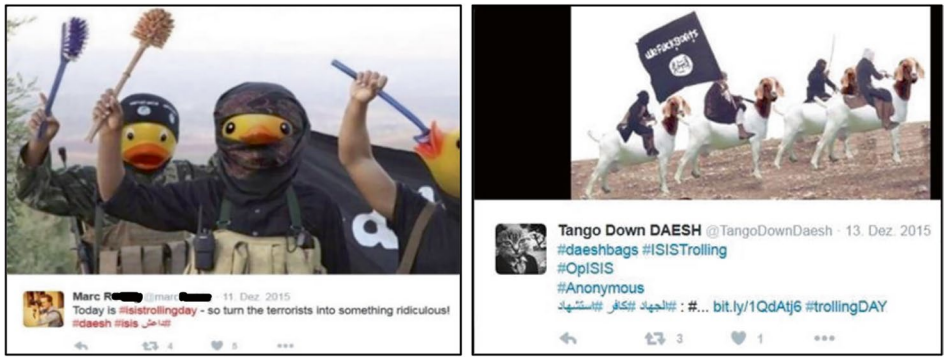

Fig. 18.6 Parody and satire used with the hashtag "#TrollingDay", showing ISIS fighters as rubber ducks and riding on goats. Source: Reuter et al. (2017)

> to adjust counterterrorism strategies to the new arenas, applying new types of measures including intelligence gathering, applying new counter measures, and training law enforcement officers specializing in the cyber domain. Researchers [from various disciplines] are coming together to develop tools and techniques to respond to terrorism's online activity. (2016, p. 89–90)

As a long-term strategy to combat radicalisation and recruitment, Weimann (2016) adds the construction of counter-narratives. Yet, (believable) anti-propaganda does not only come from the outside: Under the heading of "Anti-IS Humor", Al-Rawi (2016) explains that hundreds of Arabic YouTubers began to transform an ISIS video with religious singing into a funny dance clip after its release. In this way, parody and satire are used to mock ISIS fighters (see Fig. 18.6).

Borelli (2023) emphasises the part played by major tech corporations, including Google, Facebook and Twitter/X in countering terrorism on the internet. It is noted that these firms are shifting from a reactive to a more proactive approach in tackling this issue. Moreover, Borelli (2023) outlines four principal areas of major tech corporations' participation: policymaking, content moderation, human resources and private multilateralism. However, it is important to consider the potential impact on freedom of expression that may result from adopting a more proactive approach.

18.6 Discussion and Conclusion

In this chapter, we examined three phenomena, fake news, hate speech and online terrorism recruitment, that take place in social media (Kaplan & Haenlein, 2010) where human and machine interventions potentially inflict cultural violence (Galtung, 2007). Furthermore, to prevent a negative impact of these phenomena, various countermeasures are applied, which potentially improve cultural peace in social media. A differen-

Table 18.6 Preliminary results on actors and intentions for cultural violence and peace. (Source: Own depiction)

		Actor	
		Human	**Machine**
Intention	**Malicious interventions**	Cyber bullying, fake news, hate speech, propaganda, recruitment	Account hijacking, astroturfing, fake accounts, fake posts, spam
	Positive interventions	Gatekeeping, media literacy, laws, clarification, parody/satire, hacking, counter-narratives	Crowdsourcing, detection algorithms, visual analytics

tiation of actors and intentions is provided in Tab 18.6. In terms of (manual) **human interventions**, we see that fabricated, misinterpreted and manipulated content, as well as propaganda and terrorist recruitment may inflict cultural or direct violence. Here, countermeasures are similar and include gatekeeping, media literacy and laws, as well as clarification, parody/satire and hacking. Further research could examine how often neglected actors, such as users contributing to crowdsourcing, moderators and IT-related civil society groups, can contribute to solutions, bringing together IT knowledge and society-level interventions. These can be inspired by established peace interventions from other domains, such as reconciliation. For instance, tailored social media guidelines could improve journalistic processes or increase the population's media literacy (Kaufhold et al., 2019).

Considering (semi-)automatic **machine interventions**, we identified account hijacking, astroturfing, fake accounts, fake posts and spam as potentials for cultural violence exacerbating existing divides and eroding trust in legitimate protest and institutions. Respective countermeasures contain crowdsourcing, detection algorithms and visual analytics for malicious content. Experiences in countering spam show the power of technical arms races (Yang et al., 2019), but also spammers' adaptability in using sophisticated social engineering to deceive detection mechanisms and humans by exploiting trust detection mechanisms. Similarly, the Russian bot company IRA had adapted its strategy of feigning affiliation with established, trusted institutions (Newman, 2020), before it was disbanded due to a conflict between Russian President Putin and the IRA's founder and head of the Wagner Group. Technical arms races can thus be powerful, but never all-encompassing, leaving the necessity for social interventions. Hybrid forms of intervention include solutions that, without outright censoring posts, limit the visibility or spreading speed of harmful content, provide technical assistance for users to better judge the trustworthiness of online information, or identify social media users at risk of radicalisation.

The research field of social media analytics contributes important insights regarding cultural interventions. It deals with methods of analysing social media data and com-

prises the steps of discovery, collection, preparation and analysis (Stieglitz, Mirbabaie, Ross, et al., 2018). Current methods of social media analytics are primarily driven by domains such as businesses, crisis communication, as well as journalism and political communication (see Chap. 19 "*Political Activism on Social Media in Conflict and War*"). Social media analytics can be used to better understand the social side of social media abuse, e.g. by making situational assessments of specific discourses and events, including the identification of fake news or hate speech as potential instances of cultural violence using (supervised) machine learning approaches (Kaufhold, 2021). As an intermediary, technical tools can be developed to flag false content and provide transparency over actors and organisations that fuel the extremes and follow partisan interests. This will require identifying the actors and incentive structures that motivate disinformation and the buying of social bot systems as well as addressing the societal structures, mainly mistrust and grievances, which allow malicious interventions to take devastating effects. Although these areas have a potential impact on cultural violence and peace, it seems worthwhile examining the potentials of social media analytics and its methods for cultural peace in social media by allowing situational assessments in everyday life or during specific discourses and events (Vieweg et al., 2010).

18.7 Exercises

Exercise 18-1: What are the definitions and relations between direct, structural and cultural violence?
Exercise 18-2: What are human cultural interventions in social media? Give two examples for each negative and positive interventions and describe them briefly.
Exercise 18-3: What are automatic cultural interventions in social media? Give two examples for each negative and positive interventions and describe them briefly.
Exercise 18-4: Are automatic and human cultural interventions inherently disjoint or can they be applied in combination? Please discuss at least two examples supporting your reasoning.
Exercise 18-5: What are differences and commonalities when comparing interventions against online fake news and hate speech? Explain three aspects each.
Exercise 18-6: What countermeasures are there to prevent terrorist propaganda and recruitment in social media? Is censorship useful in this context?
Exercise 18-7: Aspects such as political activism, fake news detection, counterterrorism, and social bot detection are discussed in the light of positive cultural interventions. However, can they also exert cultural violence? Please justify your answer and give examples for at least two categories.
Exercise 18-8: Aspects such as political activism, fake news detection, counterterrorism, social bot detection as well as chat, news and warning bots are discussed in the light of positive cultural interventions. However, can they also exert cultural violence? Please justify your answer and give examples for at least two categories.

References

Recommended Reading

Reuter, C., Hartwig, K., Kirchner, J., & Schlegel, N. (2019). Fake News Perception in Germany: A Representative Study of People's Attitudes and Approaches to Counteract Disinformation. In *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*. Siegen.

Alfano, M., Carter, J., & Cheong, M. (2018). Technological Seduction and Self-Radicalization. *Journal of the American Philosophical Association*. *4*(3), 298–322. https://doi.org/10.1017/apa.2018.27.

Stieglitz, S., Brachten, F., Ross, B., & Jung, A.-K. (2017). Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts. *Proceedings of the Australasian Conference on Information Systems*, 1–11.

Bibliography

Abokhodair, N., Yoo, D., & McDonald, D. W. (2015). Dissecting a Social Botnet. *Proceedings of the Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, 839–851. https://doi.org/10.1145/2675133.2675208.

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, *13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5.

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211.

Allen, J., Martel, C., & Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3502040.

Almaatouq, A., Shmueli, E., Nouh, M., Alabdulkareem, A., Singh, V. K., Alsaleh, M., Alarifi, A., Alfaris, A., & Pentland, A. (2016). If it looks like a spammer and behaves like a spammer, it must be a spammer: Analysis and detection of microblogging spam accounts. *International Journal of Information Security*, *15*(5), 475–491. https://doi.org/10.1007/s10207-016-0321-5.

Al-Rawi, A. (2016). Anti-ISIS Humor: Cultural Resistance of Radical Ideology. *Politics, Religion & Ideology*, *7689*(May), 1–17. https://doi.org/10.1080/21567689.2016.1157076.

Al-Saggaf, Y. (2016). Understanding Online Radicalisation Using Data Science. *International Journal of Cyber Warfare and Terrorism (IJCWT)*, *6*(4), 13–27. https://doi.org/10.4018/IJCWT.2016100102.

Avery, I. (2021, Januar 20). Talkin' Bout A Revolution: Four Reasons Why the Term 'Arab Spring' is Still Problematic. *Middle East Centre, London School of Economics*. https://blogs.lse.ac.uk/mec/2021/01/20/talkin-bout-a-revolution-four-reasons-why-the-term-arab-spring-is-still-problematic/.

Awan, I. (2017). Cyber-Extremism: Isis and the Power of Social Media. *Society*, *54*(2), 138–149. https://doi.org/10.1007/s12115-017-0114-0.

Banaji, S., & Bhat, R. (2021). *Social Media and Hate* (1. Aufl.). Routledge. https://doi.org/10.4324/9781003083078.

Barzilai-Nahon, K. (2009). Gatekeeping: A critical review. *Annual Review of Information Science and Technology*, *43*(1), 1–79. https://doi.org/10.1002/aris.2009.1440430117.

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*. https://doi.org/10.1145/3544558.

Becker, B. W. (2016). The Librarian's Information War. *Behavioral & Social Sciences Librarian*, *35*(4), 188–191. https://doi.org/10.1080/01639269.2016.1284525.

Behzad, B., Bheem, B., Elizondo, D., & Martonosi, S. (2023). Prevalence and Propagation of Fake News. *Statistics and Public Policy*, *10*(1), 2190368. https://doi.org/10.1080/2330443X.2023.2190368.

Beitzinger, F., & Leest, U. (2021). *Mobbing und Cybermobbing bei Erwachsenen: Eine empirische Bestandsaufnahme in Deutschland, Österreich und der deutschsprachigen Schweiz*.

BMFSFJ. (2022). *Was ist Cybermobbing?* https://www.bmfsfj.de/bmfsfj/themen/kinder-und-jugend/medienkompetenz/was-ist-cybermobbing--86484.

Booth, K. (2007). *Theory of World Security*. Cambridge University Press.

Borelli, M. (2023). Social media corporations as actors of counter-terrorism. *New Media & Society*, *25*(11), 2877–2897.

Bunde, E. (2021). AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1264–1273. https://aisel.aisnet.org/hicss-54/da/xai/2/.

Bundeszentrale für politische Bildung. (2020, November 26). *Antisemitismus im Internet und den sozialen Medien*. bpb.de. https://www.bpb.de/themen/antisemitismus/dossier-antisemitismus/321584/antisemitismus-im-internet-und-den-sozialen-medien/.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608.

Chen, J., Yang, Z., & Yang, D. (2020). MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. *arXiv*. https://doi.org/10.18653/v1/2020.acl-main.194.

Chenzi, V. (2021). Fake news, social media and xenophobia in South Africa. *African Identities*, *19*(4), 502–521. https://doi.org/10.1080/14725843.2020.1804321.

Cheong, M., & Lee, V. C. S. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers*, *13*(1), 45–59. https://doi.org/10.1007/s10796-010-9273-x.

Chiu, K.-L., & Alexander, R. (2021). Detecting Hate Speech with GPT-3. *arXiv*.

Cho, C. H., Martens, M. L., Kim, H., Rodrigue, M., Journal, S., December, N., Kim, H., & Rodrigue, M. (2011). Astroturfing Global Warming: It Isn't Always Greener on the Other Side of the Fence. *Journal of Business Ethics*, *104*(4), 571–587. https://doi.org/10.1007/s10551-011-0950-6.

Chorev, H. (2019). Palestinian Social Media and Lone-Wolf Attacks: Subculture, Legitimization, and Epidemic. *Terrorism and Political Violence*, *31*(6), 1284–1306. https://doi.org/10.1080/09546553.2017.1341878.

Christoph, S. (2015). Funktionslogik terroristischer Propaganda im bewegten Bild. *Journal for Deradicalization*, *Fall/15*(4), 145–205.

Citron, D. K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, *91*, 1435.

Cooke, N. A. (2017). Posttruth, Truthiness, and Alternative Facts: Information Behavior and Critical Information Consumption for a New Age. *The Library Quarterly*, *87*(3), 211–221. https://doi.org/10.1086/692298.

Das NETTZ gGmbH. (2023). *Vernetzungsstelle gegen Hate Speech*. https://www.das-nettz.de/.

Deutsche Welle. (2022, Februar 27). *Fünf Fakes vom Ukraine-Krieg*. Deutsche Welle. https://www.dw.com/de/faktencheck-video-f%C3%BCnf-fakes-vom-ukraine-krieg/video-60934274.

Doncel-Martín, I., Catalan-Matamoros, D., & Elías, C. (2023). Corporate social responsibility and public diplomacy as formulas to reduce hate speech on social media in the fake news era. *Corporate Communications: An International Journal*, *28*(2), 340–352. https://doi.org/10.1108/CCIJ-04-2022-0040.

ECRI. (2015). *ECRI General Policy Recommendation N°15*. https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15.

Eismann, K., Posegga, O., & Fischbach, K. (2018). Decision Making in Emergency Management: The Role of Social Media. *Proceedings of the 26th European Conference on Information Systems (ECIS)*, 1–20.

Elmas, T., Overdorf, R., Ozkalay, A. F., & Aberer, K. (2021). Ephemeral Astroturfing Attacks: The Case of Fake Twitter Trends. *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 403–422. https://doi.org/10.1109/EuroSP51992.2021.00035.

Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection. *IEEE Access*, *9*, 103541–103563. https://doi.org/10.1109/ACCESS.2021.3098979.

Euronews. (2022, August 31). *Die 5 Top Fake News über den Ukraine-Krieg*. euronews. https://de.euronews.com/my-europe/2022/08/31/die-5-top-fake-news-uber-den-ukraine-krieg.

Faktenfinder. (2017). *AfD spokesman Christian Lüth in an interview with Faktenfinder*. http://faktenfinder.tagesschau.de/inland/falsches-antifa-foto-101.html.

Feng, S., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. *59t Annual Meeting of the Association for Computational Linguistcs and the 10th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4). https://doi.org/10.1145/3232676.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.14991.

Freedman, L. (2006). *The Transformation of Strategic Affairs*. Routledge.

Fuchs, C. (2021). How did Donald Trump incite a coup attempt? TripleC: Communication, Capitalism & Critique. *Open Access Journal for a Global Sustainable Information Society, 19*(1), 246–251.

Galtung, J. (2007). *Frieden mit friedlichen Mitteln. Friede und Konflikt, Entwicklung und Kultur*. Agenda Verlag.

Gartenstein-Ross, D. (2015). Social Media in the Next Evolution of Terrorist Recruitment. *Hearing before the Senate Committee on Homeland Security & Governmental Affairs, Foundation for Defense of Democracies*, 1–11.

Gates, S., & Podder, S. (2015). Social Media, Recruitment, Allegiance and the Islamic State. *Perspectives on Terrorism*, *9*(4), 107–116.

Gencoglu, O. (2021). Cyberbullying Detection With Fairness Constraints. *IEEE Internet Computing*, *25*(1), 20–29. https://doi.org/10.1109/MIC.2020.3032461.

Georgiou, M., & Zaborowski, R. (2017). *Media coverage of the "refugee crisis": A cross-European perspective* (DG1(2017)03). Council of Europe.

Ghanem, B., Rosso, P., & Rangel, F. (2020). An Emotional Analysis of False Information in Social Media and News Articles. *ACM Transactions on Internet Technology*, *20*(2), 1–18. https://doi.org/10.1145/3381750.

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2021). Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety*, *1*(1), Article 1. https://doi.org/10.54501/jots.v1i1.15.

Godulla, A., Hoffmann, C. P., & Seibert, D. (2021). Dealing with deepfakes – an interdisciplinary examination of the state of research and implications for communication studies. *SCM Studies in Communication and Media*, *10*(1), 72–96. https://doi.org/10.5771/2192-4007-2021-1-72.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Graves, L. (2018). Boundaries Not Drawn: Mapping the institutional roots of the global fact-checking movement. *Journalism Studies*, *19*(5), 613–631. https://doi.org/10.1080/1461670X.2016.1196602.

Gregory, S. (2022). Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism*, *23*(3), 708–729. https://doi.org/10.1177/14648849211060644.

Gundelach, P. (1979). Grass Roots Organizations. *Acta Sociologica*, *22*(2), 187–189. https://doi.org/10.1177/000169937902200206.

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). *Tweetcred: Real-time credibility assessment of content on twitter*. 228–243.

Hamid, N., & Ariza, C. (2022). *Offline Versus Online Radicalisation: Which is the Bigger Threat? Tracing Outcomes of 439 Jihadist Terrorists Between 2014–2021 in 8 Western Countries* (Global Network on Extremism and Technology (GNET)). King's Collecge, University London.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, *45*(1), 1–23. https://doi.org/10.1080/01638530701739181.

Hartwig, K., & Reuter, C. (2019). TrustyTweet: An Indicator-based Browser-Plugin to Assist Users in Dealing with Fake News on Twitter. *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*.

Hassmelden. (2022). *Melde Hatespeech. Unterstütze Betroffene. Sorge für Strafverfolgung. Verpflichte die Politik.* https://hassmelden.de/.

Haunschild, J., Kaufhold, M.-A., & Reuter, C. (2020). Sticking with Landlines? Citizens' and Police Social Media Use and Expectation During Emergencies. *Proceedings of the International Conference on Wirtschaftsinformatik (WI) (Best Paper Social Impact Award)*, 1–16. https://doi.org/10.30844/wi_2020_o2-haunschild.

Hern, A. (2020, April 7). WhatsApp to impose new limit on forwarding to fight fake news. *The Guardian*. https://www.theguardian.com/technology/2020/apr/07/whatsapp-to-impose-new-limit-on-forwarding-to-fight-fake-news.

HMdIS. (2022). *Hessen gegen Hetze*. https://hessengegenhetze.de/node/59.

Honari, A., & Alinejad, D. (2022). Online Performance of Civic Participation: What Bot-like Activity in the Persian Language Twittersphere Reveals About Political Manipulation Mechanisms. *Television & New Media*, *23*(8), 917–938. https://doi.org/10.1177/15274764211055712.

Humprecht, E. (2019). Where 'fake news' flourishes: A comparison across four Western democracies. *Information, Communication & Society*, *22*(13), 1973–1988. https://doi.org/10.1080/1369118X.2018.1474241.

Hussain, G., & Saltman, E. M. (2014). *Jihad Trending: A Comprehensive Analysis of Online Extremism and How to Counter it*. Quilliam.

Iginio, G., Danit, G., Thiago, A., & Gabriela, M. (2015). *Countering online hate speech*. UNESCO Publishing.

Jeberson, W., & Sharma, L. (2015). Survey on counter Web Terrorism. *COMPUSOFT, An international journal of advanced computer technology*, *4*(5), 1744–1747.

Jiang, M., Cui, P., Beutel, A., Faloutsos, C., & Yang, S. (2016). Catching Synchronized Behaviors in Large Networks: A Graph Mining Approach. *ACM Trans. Knowl. Discov. Data*, *10*(4), 35:1---35:27. https://doi.org/10.1145/2746403.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003.

Kaufhold, M.-A. (2021). *Information Refinement Technologies for Crisis Informatics: User Expectations and Design Principles for Social Media and Mobile Apps*. Springer Vieweg. https://doi.org/10.1007/978-3-658-33341-6.

Kaufhold, M.-A., Bayer, M., Bäumler, J., Reuter, C., Mirbabaie, M., Stieglitz, S., Basyurt, A. S., Fuchß, C., & Eyilmez, K. (2023). CYLENCE: Strategies and Tools for Cross-Media Reporting, Detection, and Treatment of Cyberbullying and Hatespeech in Law Enforcement Agencies. . . *September*.

Kaufhold, M.-A., Bayer, M., & Reuter, C. (2020). Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management*, *57*(1), 1–32. https://doi.org/10.1016/j.ipm.2019.102132.

Kaufhold, M.-A., Gizikis, A., Reuter, C., Habdank, M., & Grinko, M. (2019). Avoiding Chaotic Use of Social Media during Emergencies: Evaluation of Citizens' Guidelines. *Journal of Contingencies and Crisis Management (JCCM)*, 1–16. https://doi.org/10.1111/1468-5973.12249.

Kaufhold, M.-A., & Reuter, C. (2019). Cultural Violence and Peace in Social Media. In C. Reuter (Hrsg.), *Information Technology for Peace and Security—IT-Applications and Infrastructures in Conflicts, Crises, War, and Peace* (P. 361–381). Springer Vieweg. https://doi.org/10.1007/978-3-658-25652-4_17.

Kaufhold, M.-A., Rupp, N., Reuter, C., & Habdank, M. (2020). Mitigating Information Overload in Social Media during Conflicts and Crises: Design and Evaluation of a Cross-Platform Alerting System. *Behaviour & Information Technology (BIT)*, *39*(3), 319–342. https://doi.org/10.1080/0144929X.2019.1620334.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Hrsg.), *Information Visualization* (Bd. 4950, pp. 154–175). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7

Khaldarova, I., & Pantti, M. (2020). Fake news: The narrative battle over the Ukrainian conflict. In *The Future of Journalism: Risks, Threats and Opportunities* (P. 228–238). Routledge.

Klausen, J., Barbieri, E. T., Reichlin-Melnick, A., & Zelin, A. Y. (2012). The YouTube Jihadists: A Social Network Analysis of Al-Muhajiroun's Propaganda Campaign. *Perspectives on Terrorism*, *6*(1), 36–53.

Klein, D. O., & Wueller, J. R. (2017). Fake news: A legal perspective. *Journal Of Internet Law*, *20*(10), 6–13.

Kouvela, M., Dimitriadis, I., & Vakali, A. (2020). Bot-Detective: An explainable Twitter bot detection service with crowdsourcing functionalities. *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, 55–63. https://doi.org/10.1145/3415958.3433075.

Landesanstalt für Medien NRW. (2021). *Forsa-Befragung zur Wahrnehmung von Hassrede*.

López-Martínez, A., García-Díaz, J. A., Valencia-García, R., & Ruiz-Martínez, A. (2019). CyberDect. A novel approach for cyberbullying detection on twitter. *International Conference on Technologies and Innovation*, 109–121.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*.

Maniou, T. A., & Veglis, A. (2020). Employing a Chatbot for News Dissemination during Crisis: Design , Implementation and Evaluation. *future internet Article*, *12*(109), 1–14.

McCarthy, T. (2017). How Russia used social media to divide Americans. *The Guardian*. https://www.theguardian.com/us-news/2017/oct/14/russia-us-politics-social-media-facebook.

Mihailidis, P., & Viotty, S. (2017). Spreadable Spectacle in Digital Culture: Civic Expression, Fake News, and the Role of Media Literacies in "Post-Fact" Society. *American Behavioral Scientist*, *61*(4), 441–454. https://doi.org/10.1177/0002764217701217.

Miró-Llinares, F., & Aguerri, J. C. (2023). Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'. *European Journal of Criminology*, *20*(1), 356–374.

Mohan, S. (2015). Locating the "Internet Hindu": Political Speech and Performance in Indian Cyberspace. *Television & New Media*, *16*(4), 339–345. https://doi.org/10.1177/1527476415575491.

Mostrous, A., Bridge, M., & Gibbons, K. (2017). *Russia used Twitter bots and trolls 'to disrupt' Brexit vote*. https://www.thetimes.co.uk/article/russia-used-web-posts-to-disrupt-brexit-vote-h9nv5zg6c.

Mou, G., & Lee, K. (2021). An Effective, Robust and Fairness-aware Hate Speech Detection Framework. *IEEE International Conference on Big Data*, 687–697. https://doi.org/10.1109/bigdata52589.2021.9672022.

Mozur, P. (2018, Oktober 15). A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times*. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

Müller, P. Dr., & Denner, N. (2017). *Was tun gegen „Fake News"?*

Nanath, K., Kaitheri, S., Malik, S., & Mustafa, S. (2022). Examination of Fake News from a Viral Perspective: An Interplay of Emotions, Resonance, and Sentiments. *Journal of Systems and Information Technology*, *24*(2), 131–155. https://doi.org/10.1108/JSIT-11-2020-0257.

Narwal, V., Salih, M. H., Lopez, J. A., Ortega, A., O'Donovan, J., Höllerer, T., & Savage, S. (2017). Automated Assistants to Identify and Prompt Action on Visual News Bias. *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2796–2801. https://doi.org/10.1145/3027063.3053227.

Neue Wege des Lernens e.V. (2017). *Fake News Check*. https://www.neue-wege-des-lernens.de/2017/03/19/fake-news-check-mit-dem-smartphone/.

Newman, L. H. (2020). Russia Is Learning How to Bypass Facebook's Disinfo Defenses. *Wired*. https://www.wired.com/story/russia-ira-bypass-facebook-disinfo-defenses/.

Ka Chung Ng, Jie Tang & Dongwon Lee (2021) The Effect of Platform Intervention Policies on Fake News Dissemination and Survival: An EmpiricalExamination. *Journal of Management Information Systems, 38*(4), 898–930, https://doi.org/10.1080/07421222.2021.1990612

Oboler, A. (2014). *The antisemitic meme of the Jew*. Online Hate Prevention Institute.

Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing Big Social Data. *Journal of Big Data*, *4*(1), 1–19. https://doi.org/10.1186/s40537-017-0063-x.

Onorati, T., Díaz, P., & Carrion, B. (2019). From social networks to emergency operation centers: A semantic visualization approach. *Future Generation Computer Systems*, *95*, 829–840. https://doi.org/10.1016/j.future.2018.01.052.

Paschalides, D., Stephanidis, D., Andreou, A., Orphanou, K., Pallis, G., Dikaiakos, M. D., & Markatos, E. (2020). Mandola: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Online Hate Speech. *ACM Transactions on Internet Technology*, *20*(2), 1–21. https://doi.org/10.1145/3371276.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, *55*(2), 477–523. https://doi.org/10.1007/s10579-020-09502-8.

Potter, W. J. (2010). The state of media literacy. *Journal of Broadcasting and Electronic Media*, *54*(4), 675–696. https://doi.org/10.1080/08838151.2011.521462.

Reuter, C. (2020). Towards IT Peace Research: Challenges at the Intersection of Peace and Conflict Research and Computer Science. *S+F Sicherheit und Frieden / Peace and Security*, *38*(1), 10–16. http://www.peasec.de/paper/2020/2020_Reuter_TowardsITPeaceResearch_SF.pdf. https://doi.org/10.5771/0175-274X-2020-1-10.

Reuter, C., Hartwig, K., Kirchner, J., & Schlegel, N. (2019). Fake News Perception in Germany: A Representative Study of People's Attitudes and Approaches to Counteract Disinformation. *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*.

Reuter, C., & Kaufhold, M.-A. (2018). Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics. *Journal of Contingencies and Crisis Management (JCCM)*, *26*, 1–17.

Reuter, C., Pätsch, K., & Runft, E. (2017). IT for Peace? Fighting Against Terrorism in Social Media – An Explorative Twitter Study. *i-com: Journal of Interactive Media*, *16*(2), 181–195.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). „Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

Rizos, G., Hemker, K., & Schuller, B. (2019). Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. *International Conference on Information and Knowledge Management (CIKM)*. https://doi.org/10.1145/3357384.3358040.

Robinson, T., Callahan, C., Boyle, K., Rivera, E., & Cho, J. K. (2017). I like FB: A Q-Methodology Analysis of Why People 'Like' Facebook. *International Journal of Virtual Communities and Social Networking (IJVCSN)*, *9*(2), 46–61. https://doi.org/10.4018/IJVCSN.2017040103.

Rossoliński-Liebe, G., & Willems, B. (2022). Putin's Abuse of History: Ukrainian 'Nazis','Genocide', and a Fake Threat Scenario. *The Journal of Slavic Military Studies*, *35*(1), 1–10.

Rubin, V. L. (2019). Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions. *Journal of Documentation*, *75*(5), 1013–1034. https://doi.org/10.1108/JD-12-2018-0209.

Saltman, E. M., & Russell, J. (2014). *White Paper – The role of prevent in countering online extremism*. Quilliam.

Sängerlaub, A. (2017a). *Deutschland vor der Bundestagswahl: Überall Fake News?!* Stiftung Neue Verantwortung.

Sängerlaub, A. (2017b). *Verzerrte Realitäten: „Fake News" im Schatten der USA und der Bundestagswahl*. Stiftung Neue Verantwortung.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2020). The risk of racial bias in hate speech detection. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 1668–1678. https://doi.org/10.18653/v1/p19-1163.

Schoch, D., Keller, F. B., Stier, S., & Yang, J. (2022). Coordination patterns reveal online political astroturfing across the world. *Scientific Reports*, *12*(1), 4572. https://doi.org/10.1038/s41598-022-08404-9.

Shi, H., Livescu, K., & Gimpel, K. (2021). Substructure Substitution: Structured Data Augmentation for NLP. *arXiv*.

Soden, R., & Palen, L. (2018). Informating Crisis: Expanding Critical Perspectives in Crisis Informatics. *Proceedings of the ACM on Human-Computer Interaction*.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv*.

Stern. (2023). *„Arma 3": Fake-Video soll Nahostkonflikt zeigen*. https://www.stern.de/digital/web-video/fake-or-no-fake/-arma-3---fake-video-soll-nahostkonflikt-zeigen--video--30530564.html.

Stieglitz, S., Brachten, F., Ross, B., & Jung, A.-K. (2017). Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts. *Proceedings of the Australasian Conference on Infor-mation Systems*, 1–11.

Stieglitz, S., Hofeditz, L., Brünker, F., Ehnis, C., Mirbabaie, M., & Ross, B. (2022). Design principles for conversational agents to support Emergency Management Agencies. *International Journal of Information Management*, *63*. https://doi.org/10.1016/j.ijinfomgt.2021.102469.

Stieglitz, S., Mirbabaie, M., & Milde, M. (2018). Social Positions and Collective Sense-Making in Crisis Communication. *International Journal of Human–Computer Interaction*, *34*(4), 328–355. https://doi.org/10.1080/10447318.2018.1427830.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002.

Tagesschau. (2018). *Propaganda in Syrien: Zwischen Fiktion und Wirklichkeit*. tagesschau.de. https://www.tagesschau.de/faktenfinder/fake-syrien-revolutionman-101.html.

Tagesschau. (2023). *Angriff auf Israel: Zahlreiche Falschmeldungen kursieren im Netz*. tagesschau.de. https://www.tagesschau.de/faktenfinder/israel-hamas-fakes-100.html.

Torok, R. (2015). *ISIS and the Institution of Online Terrorist Recruitment*. Middle East Institute. https://www.mei.edu/publications/isis-and-institution-online-terrorist-recruitment.

Trang, D., Johansson, F., & Rosell, M. (2015). Evaluating Algorithms for Detection of Compromised Social Media User Accounts. *Proceedings - 2nd European Network Intelligence Conference, ENIC 2015*, 75–82. https://doi.org/10.1109/ENIC.2015.19.

United States Holocaust Memorial Museum. (2023). *Nazi-era Antisemitic Propaganda Poster*. https://encyclopedia.ushmm.org/content/en/photo/anti-jewish-propaganda.

Verstraete, M., Bambauer, J. R., & Bambauer, D. E. (2022). Identifying and countering fake news. *Hastings LJ*, *73*, 821.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. *In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, 1079–1088. https://doi.org/10.1145/1753326.1753486.

Viviani, M., & Pasi, G. (2017). Credibility in social media: Opinions, news, and health information—A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(5), e1209--n/a. https://doi.org/10.1002/widm.1209.

Volkova, S., & Jang, J. Y. (2018). Misleading or Falsification: Inferring Deceptive Strategies and Types in Online News and Social Media. *Companion Proceedings of the The Web Conference 2018*, 575–583. https://doi.org/10.1145/3184558.3188728.

Waever, O. (1993). Societal security: The concept. *Identity, migration and the new security agenda in Europe*, 17–40.

Webel, C., & Galtung, J. (2007). Negotiation and international conflict. In *Handbook of Peace and Conflict* (Nummer 11881, P. 35–50). Routledge. https://doi.org/10.4324/9780203089163.ch3.

Weimann, G. (2016). The Emerging Role of Social Media in the Recruitment of Foreign Fighters. In A. de Guttry, F. Capone, & C. Paulussen (Hrsg.), *Foreign Fighters under International Law and Beyond*, 77–95. T.M.C. Asser Press. https://doi.org/10.1007/978-94-6265-099-2_6.

Weimann, G., & Jost, J. (2015). Neuer Terrorismus und Neue Medien. *Zeitschrift für Außen- und Sicherheitspolitik*, *8*(3), 369–388. https://doi.org/10.1007/s12399-015-0493-5.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review*, *9*(11).

Whittaker, J. (2022). Rethinking Online Radicalization. *Terrorism Research Initiative*, *16*(4).

Wirtschafter, V., & Majumder, S. (2023). Future Challenges for Online, Crowdsourced Content Moderation: Evidence from Twitter's Community Notes. *Journal of Online Trust and Safety*, *2*(1).

Wohn, D. Y., Fiesler, C., Hemphill, L., De Choudhury, M., & Matias, J. N. (2017). How to Handle Online Risks?: Discussing Content Curation and Moderation in Social Media. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1271–1276. https://doi.org/10.1145/3027063.3051141.

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. http://arxiv.org/abs/2210.15723.

Wu, X., Fan, W., Gao, J., Feng, Z. M., & Yu, Y. (2015). Detecting Marionette Microblog Users for Improved Information Credibility. *Journal of Computer Science and Technology*, *30*(5), 1082–1096. https://doi.org/10.1007/s11390-015-1584-4.

Yang, K., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. https://doi.org/10.1002/hbe2.115.

Yin, W., & Zubiaga, A. (2021). *Towards generalisable hate speech detection: A review on obstacles and solutions* (arXiv:2102.08886). arXiv. https://doi.org/10.48550/arXiv.2102.08886.

Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., & Starbird, K. (2018). From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW). https://doi.org/10.1145/3274464.

Ziegele, M., Breiner, T., & Quiring, O. (2014). What Creates Interactivity in Online News Discussions? An Exploratory Analysis of Discussion Factors in User Comments on News Items. *Journal of Communication*, *64*(6), 1111–1138. https://doi.org/10.1111/jcom.12123.