CHAPTER 11

Countering Disinformation: Corporations and Civil Society

Instances of significant manipulation through social media during election campaigns have been recorded since the mid-2000s. For example, Putin's staff conducted dress rehearsals before the 2007 and 2008 presidential elections. Additionally, in his book *This is Not Propaganda,* Peter Pomerantsev describes how Rodrigo Duterte used social media to win the presidential election in the Philippines in 2012 (Pomerantsev, 2019, pp. 1–16). Following the experience of social media manipulation in the 2016 U.S. presidential election, the world's democracies began gradually taking more active measures against foreign manipulators.

1 Social Platforms: Evolution and Disinformation

During the 2017 French presidential election, a joint French and British media cluster created a platform for verification, fact-checking, and debunking of counterfeits, aimed at ensuring fair conduct that reflects the will of voters (*CrossCheck France*, 2021). Similarly, prior to the federal parliamentary elections in Germany in September 2021, a national front was established to counteract disinformation. It was supported not only by the media but also research centers, including international centers, federal government institutions, and the political parties participating in the elections.

Combating disinformation is not limited to politics and elections. In the era of digital disinformation, the biggest challenge in countering it is not just individual distortions or manipulations, or even coordinated operations of influence, but the hard-to-contain nature of social media itself. Disinformation is perpetuated by the constant presence of multiple narratives, which create and maintain information chaos.

Platforms' features overlap with the power of their corporations, and the potential benefits of dividing information giants into smaller companies remain a lofty, unrealistic goal. These platforms not only develop their empires but also produce devices, monopolizing both the channels of communication and the increasing production of necessary tools.

To illustrate, Apple has the power to remove from its Apple Store any Facebook/Meta or Google applications that it deems harmful due to their potential for disseminating disinformation. Google will not do the same with applications created for its own smartphones, Pixel. Technological giants wield influence and economic opportunities on par with states. For instance, Meta's annual income is nearly $100 billion USD, equivalent to the domestic product of a large wealthy country. Their income will only continue to increase, not only from internet services, but also from investments in banking or cryptocurrency production. For Mark Zuckerberg, the company's importance takes priority over democracy, and the amount of knowledge Meta has about its users exceeds what they voluntarily provide. According to experts, an average of 70 Facebook "likes" provides more data than an individual's friends have, while 300 "likes" provides more data than spouses possess (Haden, 2021).

The accumulation and aggregation of user data by corporations, coupled with their increasing economic power, has resulted in a host of issues, such as the challenge of accessing service provider databases. Experts struggle to access these "black boxes", which makes it difficult to pinpoint potential risks associated with the use of these services and how they may mislead users. These databases gather information about users and their online behavior while also revealing the actions taken by platform managers to combat manipulation in web administration.

Effective measures to combat disinformation on social media platforms leave much to be desired, and without conscientious regulatory obligations, particularly from the largest states and international organizations, efforts will continue to fall short. The ambivalent approach of corporations and platforms such as Facebook, Google, or Twitter toward this issue is not surprising given that their business models prioritize instinctive user action and emotional involvement. While technological giants have taken steps to reduce falsehood and hate speech online, as well as potential manipulation and disinformation, decisive statements are often not followed up with sufficient action. Instead, they tend to be reactive and largely prompted by media pressure, research organizations, governments, and international institutions. Furthermore, smaller but increasingly powerful entities such as TikTok and Telegram demonstrate passivity in addressing the issue. The following examples illustrate the approach used by technology companies toward reducing falsehood and aggression online.

1.1 *Facebook*

– Facebook conducts its own research and activities to limit disinformation on the web.

– A monitoring board for web content was established. They publish periodic reports on moderation, warnings, content deletion, and account blocking, which includes part of the Code of Conduct for Countering Disinformation.
– Facebook collaborates in fact-checking and training for journalists and covers the cost of the trainings; the platform therapeutically redirects users from risk groups to sources advising on how to counteract extremism, which is also known as the "redirect initiative".
– It conducts tests to assess the limitation of visibility of political content in the United States, Canada, Brazil, and Indonesia. Tests are also conducted for functions like "read before posting".
– Moderators remove or suspend accounts, including those of politicians who repeatedly spread disinformation or extremist views.
– The Taliban is banned from being active on the platform.

1.2 *YouTube/Google*

– Both YouTube and Google conduct research on disinformation and methods of counteracting it (Jigsaw).
– They block the share of ad revenue from a pool of the platform (i.e., they block ads targeted at people under 18).
– They remove content published by politicians and governments that violate the platform's regulations such as misinformation on health matters or content to incite hatred.

1.3 *Twitter*

– The "Birdwatch" program in the U.S., South Korea, and Australia was launched, which allows volunteers to tag tweets that contain false information.
– Twitter started the practice of pre-bunking which is pre-empting disinformation regarding internet voting.
– Accounts of politicians misinforming about COVID-19 or spreading conspiracy theories are banned.

Reddit has blocked subgroups that disseminate COVID-19 misinformation, while other platforms have taken measures such as conducting research to counteract disinformation by creating fact-checking tools, working together with media and journalists, blocking and removing accounts (including those belonging to politicians), or applying interstitial warnings.

The conclusions resulting from the information presented above show that: (1) in the face of public pressure, proactive undertakings by platforms are dominated by protecting children and young people against disinformation and aggression as well as public health-related topics and combating extremism;

(2) there has been a noticeably more active reaction to political manipulation and extremism, particularly from the side of the largest corporations; (3) these corporations have begun to limit the possibilities of earning money from disinformation.

Companies conduct their own research on disinformation and introduce and provide free research tools in this area. They also experiment and expand their cooperation with journalists. Although these tools are insufficient to fully contain the current level of disinformation, fact-checking would be much more difficult without them.

Corporations responded swiftly to Russia's invasion of Ukraine, although their implementation of certain decisions lacked consistency. YouTube and Facebook restricted access to Russian state media in the European Union, prompting Russia to retaliate by blocking Facebook and restricting Twitter's operations within the country. Russia also cut off the broadcasting of several independent Russian-language Western media outlets, including the *BBC* and *Deutsche Welle*. Microsoft has blocked Russians from purchasing its products and services, but existing customers in Russia are still able to use them. TikTok has disabled the publishing of new posts in Russia and blocked *RT* and *Sputnik*, but other regime media outlets were left untouched. Twitter faced criticism for its half-measures, as it only limited the visibility of tweets from the regime's media in Russia while remaining open to pro-Kremlin disinformation channels.

Moderation remains the primary method for removing undesirable content on a global scale. It involves managing and administering the behavior of social media users in compliance with regulations, which include rules and guidelines that limit or block the publication of unacceptable content. Moderation may occur before publication as preventive moderation or after publication, conducted algorithmically, automatically, or by a human moderator. It may also involve temporary account suspensions or permanent account deletions from a platform's community.

Online platforms typically use mixed moderation models. In its moderation policy, Google blocks ads on YouTube targeted at users under 18 based on their age, gender, or interests. The company plans to expand the scope of blocked ads by applying additional protective filters for this category of users. Additionally, Google has introduced measures to protect the security of personal data at the request of users or their guardians, such as blocking the public ability to find images of minors through its search engine. These measures aim to extend privacy and prevent advertisements that directly prompt young people to spend money. Google is adapting to the requirements imposed on platforms by lawmakers in EU and other countries. Facebook has announced similar actions.

In one particular case, YouTube removed 15 videos of Brazilian President Jair Bolsonaro due to their COVID-19 content. Bolsonaro reacted by submitting a draft law to impede similar actions in the future, but the court blocked its entry into force. Facebook also removed accounts associated with the Russian company Fazze, which conducted anti-vaccine disinformation by attempting to bribe influential Western bloggers to criticize Western vaccines and promote Russian ones. Pakistani and Bangladeshi accounts were also used in the Fazze operation. Linguistic errors in the repeated messages exposed this peculiar laundering of disinformation.

As a result of the COVID-19 pandemic, there has been a shift in the attitudes and commitments of social media platforms in combating disinformation, leading to increased collaboration with civil society groups. Facebook has funded fact-checking training for journalists, investing around $85 million between 2016 and 2021. A powerful way to discourage counterfeiters is to demonetize their content by refusing to share in the profits of advertisements placed alongside extremist, deceptive, or manipulated publications. Google has demonetized websites that spread anti-vaccine theories through their AdSense program. Facebook has introduced the Redirect Initiative, which redirects users to websites and sources that counteract extremism (Clark, 2021). The platform is also experimenting with limiting political content in the U.S., Canada, Brazil, and Indonesia, and plans to expand this tool to Spain, Ireland, and Sweden. Additionally, Facebook is testing a "read-before-share" feature, similar to the one used by Twitter for some time.

Twitter has partnered with Reuters and the Associated Press to gain contextual journalistic knowledge about public and international affairs, allowing platform moderators to work more effectively and consciously. It has also launched Birdwatch, enabling volunteers to tag tweets they believe are misleading (Timmins, 2021). The program is currently being tested in the United States, South Korea, and Australia. Additionally, Twitter has implemented pre-bunking, a means of preventing disinformation by anticipating misinformation related to internet voting (Boman, 2020). Twitter actively monitors and responds to content that violates its regulations, detecting around 65% of such content (Dang & Culliford, 2021). It removes or suspends accounts of politicians who engage in activities that threaten public safety or health, such as former U.S. President Donald Trump and other members of Congress. Twitter employs around 1,500 moderators to manage content for 200 million daily users.

Due to the growing market for on-demand disinformation, platforms are removing more accounts that commercialize this practice. Perpetuating pro-Kremlin propaganda can earn several thousand U.S. dollars per month.

Meta has removed 6,000 pro-Russian accounts, parties, and groups from Facebook and Instagram since 2017. Twitter has removed 5,000 accounts since 2018, and YouTube has removed 2 million videos since 2019. After the invasion of Ukraine, these numbers soared. Many of the accounts removed were linked to the St. Petersburg-based troll farm, the Internet Research Agency. Russia has demonstrated the importance of YouTube in their disinformation activities through their nervous reaction to Google's decision to block the German-language *RT* channel. YouTube is therefore not defenseless against the pricing measures taken by the Kremlin in Russia (Dubov, 2021).

Russians are often behind the accounts operated by organizers of foreign disinformation on social media platforms, but there are also increasing numbers of accounts operated by Chinese, Arab, Iranian, and non-state actors associated with terrorist organizations. Despite efforts to counteract internal political or foreign manipulation, political advertising through web content remains one of the key problems. Decisions on how best to monitor and regulate this issue are still being made. Experts believe that efforts should focus on:

- Defining the ads.
- Maximizing transparency and verification by, for instance, determining who is behind the ad.
- Limiting the proliferation of advertisers of unknown provenance, especially those who appear just before the elections.
- Limiting the number of advertisements to better hold the authors accountable.
- Distinguishing ads from other content.
- Introducing election silence on online advertising.
- Applying the silence rule consistently, not only during election campaigns.
- Enforcing the rules and punishing violators (*Ten simple*, 2021).

Regulations in this area would enable better tracking of advertisements by or for foreign clients with disinformation motives. Drafts of legal acts that have emerged in the European Union and the United States suggest that increased transparency standards can be expected in this regard.

Greater proactivity is expected from corporations on issues such as engaging with researchers and providing them with greater insight into databases, algorithmic systems, and content moderation. There is also a need for more consistent enforcement of compliance with their own regulations. A problem is the lack of uniform standards and the reactive nature of actions taken only under pressure from governments or public opinion. For example, while Google limits the possibility of advertising to minors, Facebook does not (Culliford, 2021a). In fact, Facebook is even developing applications aimed at four-year-olds

to attract new generations. Facebook prohibits the Taliban from being active on its platform, but Twitter does not. Reddit has blocked COVID-19 disinformation but only in response to protests from other users. Facebook works together with the journalism community on fact-checking and journalist training, allocating $84 million toward it in recent years. This is only a small portion of its income, however. These online platforms also engage in activities that either harm efforts to prevent and combat disinformation or simulate such efforts due to conflicts between public interest and their business goals and practices. Harmful or simulation actions taken by these online platforms are described below.

1.4 *Facebook*

- Facebook limits human control over the effects of algorithms in terms of moderation.
- It discriminates against weaker groups of network users.
- It has created categories of users who are attracting people and increasing the popularity of the platform, and they are given more exposure than others.
- It limits moderation or does not moderate content in rarer languages.
- Restrictive conditions are imposed for cooperation with researchers.
- Genuinely favorable conditions are created for recruitment to extremist groups, creating real health problems and rejecting proposals for adequate changes.

1.5 *YouTube*

- YouTube has seen a marked increase in harmful content recommendations.
- YouTube does not take measures to redirect to websites correcting disinformation; what's worse, it allows redirection to further disinforming sources.
- Its contextual warnings are not effective.

TikTok has been found to spread disinformation about the coronavirus and only remove extremist content after users flag it. Amazon and Spotify permit shows that spread anti-vaccine propaganda. Other platforms allow advertisers to target recipients of their ads; trade in gadgets that promote extremist and misinformation organizations; enable bypassing of bans, promote intolerance, phobias, and aggression; or are ineffective in preventing banned users from returning to the platform.

The issue of inconsistency in filtering published content has come to the forefront of public debate. An example of an arbitrary approach to moderating user-posted content is the creation of a specific category on Facebook's special XCheck user list. These users are given more freedom in their content due to their media coverage and can therefore attract greater engagement from others on the web. For instance, Facebook did not take action on a post by

Brazilian footballer Neymar featuring a naked woman accusing him of rape. This lack of action symbolizes a tolerance toward many celebrities who are also influencers.

Computer algorithms designed to prevent the automatic recommendation of users to join extremist groups on Facebook have limited success. Shocking data shows that 64% of recruitment to extremist groups is done through Facebook features such as "Groups You Should Join" and "Discover". However, Facebook management has rejected proposed models of change that could reduce this type of engagement on the platform (Hao, 2021a).

As part of the redirect initiative, Facebook is testing the Hints program that helps users find information and resources when they are exposed to extremist content. The program redirects users to information about programs such as Life After Hate. A global program like this may have more success than simply removing extremist messages. Given its position in the market and financial capabilities, Facebook should allocate more resources to fight disinformation. Expectations regarding this matter should be proportional to the company's profits or the social harm caused by unconscious or willful neglect.

Facebook also faces difficulties in moderating content in less common languages, as is the case in Ethiopia, a country with a population of 100 million people and six official languages. The company primarily relies on automated moderation in this market, which is unreliable and does little to limit escalating tensions, inter-ethnic conflicts, and even violence such as rape and homicide. Furthermore, Facebook faces significant moderation issues throughout the Arab world (Scott, 2021b), as well as in non-English speaking environments in general. This includes the Italian platform, where protection against disinformation is inadequate (Steffenhagen, 2021).

Although Facebook takes measures against large disinformation sources, it only does so in a more definitive way under pressure from governments or public opinion (Darcy, 2021). External pressure from governments and civil society is clearly insufficient as Facebook users remain disproportionately opposed to vaccination compared to viewers of the conservative Fox News media outlet in the United States (Kimball, 2021). Another issue is that online media creates content specifically for its potential popularity on Facebook, tailoring it to the platform's mechanisms of popularity (Hagey & Horwitz, 2021) at the expense of reliability.

With billions of users and exponentially more publications, YouTube's removal of a million fake COVID-19 posts (Solsman, 2021) is merely symbolic. On TikTok, 80% of extremist content is removed only after user intervention. In reality, after clicking on a COVID-19 disinformation page, users are not redirected to pages that correct disinformation but to other disinformation sources

(Sweet, 2021). Even after detecting and identifying misinformation or influence operations and their perpetrators on Twitter, some of them still persist.

Platforms could enhance their efforts to prevent misinformation and disinformation not only by removing problematic content but also by providing contextual alerts to their users. Such alerts could inform users that the information they are about to read or watch may be unreliable. Instead of redirecting users to the page containing falsehoods, the alert could break the link and prompt the user to make an informed decision about whether to proceed. Contextual warnings are frequently ignored by users, but warnings that interrupt reading or viewing tend to be more effective.

Researchers at Harvard Kennedy School have noted that platforms often have a narrow understanding of the fight against disinformation, limiting their approach to fact-checking. At the same time, they ignore the broader political and cultural context of the messages. Representatives of weaker and minority groups are often the targets of xenophobic and disinformation attacks due to willful ignorance of the issues. Counteracting disinformation should therefore not only concern facts and recognizing falsehoods or hate, but also examine power structures that favor disinformation, the functioning of technology companies, state agencies, and the entire media and information environment. The combined force of these factors saturates the web with anti-racist or anti-Islamic content in the context of the fight against terrorism and promotes the superiority of the white race and nationalist attitudes. The authors postulate a multidisciplinary approach to research on falsehood and disinformation that would allow for a wider inclusion of knowledge about history, political economy, and other social sciences in contemporary research on information space and media platforms (Kuo & Marwick, 2021).

Facebook's Oversight Board has been criticized for not enforcing its own recommendations on content moderation and automation of online operations. While it has made many unprecedented declarations that are consistent with those postulated by civic groups, it does not implement them when doing so exposes the company to excessive costs. The challenge, therefore, is to find a balance between freedom of expression, the extent of interference, and ultimately, the business model of social media platforms. Critics suggest a more scrupulous implementation of the council's recommendations through increasing platform transparency and better assessing the impact of algorithms on humans. They propose setting up multi-environmental social media boards to moderate and interfere with content (Kayyali & York, 2021). This approach is partly followed by designers of new solutions within the European Union, which will be discussed further down.

The true change that is expected in social media's fight against disinformation will not be achieved solely by moderating and deleting posts or blocking accounts but rather by a fundamental shift in the nature of the platforms' operations and their business models. They must prioritize social good and their own responsibilities over profits, either by their own choice or through government regulation. While moderation discussions are undoubtedly necessary, they are becoming increasingly more of a distraction from the heart of the problem, which the platforms find convenient. By focusing on moderation, these platforms are dealing with symptoms rather than addressing the disease itself (Melford & Rogers, 2021).

Improved cooperation between the scientific community, researchers, and social media platforms could help navigate the dilemmas and paradoxes described above. However, Facebook's restrictive terms of cooperation and blocking of researchers from accessing data, including the termination of their accounts, hinder progress (*Facebook*, 2021; Kaye, 2021). The company's aggressive stance against attempts to influence its operations is illustrated by Louis Barclay of *Slate*, who noted the threat of legal action the platform levied against a scientist who created an application allowing users to delete content from their timeline (Barclay, 2021). Unequal enforcement of regulations is also a problem, with content in English being removed but kept in other languages such as French. Inconsistencies in enforcing regulations and applying internal report recommendations complicate matters further.

The new negative trends and phenomena provoked by the coronavirus pandemic are well illustrated by the case of Telegram, one of the most popular platforms among Russians. Pavel Durov created it, just like the most popular social networking site among Russians, *Vkontakte,* but he was forced to sell Telegram to Kremlin-obedient oligarchs. The platform, once banned in Russia, was also used by Russian state bodies, politicians, and officials. During the presidential elections in Moldova in 2021, it was the main platform for disinformation targeting democratic candidates.

The popularity of Telegram has grown significantly in recent years, and it has become a platform for various activities, including extremist content and disinformation campaigns. According to Pipe (2021), Telegram has become a haven for extremists due to its lack of moderation policies. Additionally, its popularity is increasing in countries such as Germany and Spanish-speaking regions (Loucaides, 2021). As of July 2021, Telegram had 550 million users, making it a significant source of mass disinformation (Talant, 2021). Due to its accessibility and popularity in Russia, however, it can also be an essential

platform for sharing accurate information about issues such as the war in Ukraine, used by Ukrainian expatriates in particular.

One unintended consequence of restrictions on open platforms is that users migrate to encrypted platforms. This leads to an increase in the popularity, income, and influence of the encrypted platforms in the market and among users.

Although corporations cannot be held responsible for all wrongdoing, it is clear that despite efforts to combat misinformation, spreading false information on social media is still profitable and even encouraged. With an annual revenue of $400 billion, Amazon could easily refuse to sell books that promote anti-vaccine ideology, while Spotify has no need to make money off podcasts that question the efficacy of vaccines. Other large companies should also avoid financing propaganda, such as ads on Belarusian state television. Additionally, corporations' claims that they do not sell personal user data to advertisers is hypocritical as it is widely known that data transfers allow advertisers to precisely identify and target recipients (*What Does*, 2021).

In addition to their failures in combating disinformation, these platforms also hold a monopoly on information. For example, in the Philippines, 96% of residents have Facebook accounts and rely on the platform for daily news and information. Meanwhile, Facebook's CEO, Mark Zuckerberg, still maintains that social media is not mass media, thereby avoiding regulations that traditional media outlets are subject to. Legal regulations enforced by states and international organizations are necessary to hold these platforms accountable. Former employees and whistleblowers, such as Frances Haugen, have revealed that the company's leadership has rejected proposals for changes aimed at protecting users' well-being, such as reducing their time spent on the platform. However, Haugen remains a loyal representative of the industry and opposes the idea of splitting up Facebook or amending U.S. media laws to hold social media platforms accountable for content. Instead, she advocates for changes in algorithm design, such as prioritizing chronological timelines and reducing emotionally charged content, which could lead to positive changes in content recommendation and user experience on these platforms (Hao, 2021a).

Some problems posed by platforms are easily identifiable and simple to deal with, such as tackling the posting of commercial ads with anti-Semitic, xenophobic, or homophobic content that promote intolerance and aggression (*Ad-funded*, 2021). However, other corporate misinformation offenses will require more complex countermeasures, more time, and more effort from both the platforms themselves and regulators.

The actions of corporations and their representatives, including former employees, should be taken seriously but also approached with caution. Their lobbying networks in Western countries and the European Union are powerful and reminiscent of how the tobacco and energy industries have protected their businesses while posing as champions of new solutions.

Global Disinformation Index experts have proposed various solutions for social media, including general and specific, voluntary or compulsory, political or legal measures. Some of these proposals have already been introduced by the European Union and its member states, but they primarily relate to the situation in the United States:

- Make platforms more accountable for spreading lies.
- Identify notorious disinformers and ban them.
- Use content moderation to raise awareness of falsehood and truth among users.
- Use awareness of falsehood and truth to collectively flag false information so that algorithms can identify lies more effectively and efficiently.
- Increase researchers' access to data.
- Offer platforms forms of limiting liability for the appearance of illegal content on them in exchange for researchers' access to aggregated data about users, their behavior, and methods of counteracting prohibited acts by platforms.
- Set up independent expert councils to review research spending on platforms.
- Demand anti-monopoly legal solutions.
- Create a global institution for cooperation and discussion on internet governance.
- Appoint statutory bodies to supervise the fulfilment of obligations.
- Offer platforms protection against liability when they fully comply with content moderation provisions.
- Split the largest companies, starting with Facebook.
- Regulate platform interoperability issues.
- Follow the example of the EU's proposal to impose stricter penalties.
- Create the possibility of running joint programs with the largest platforms to increase innovation (Decker & Boucher, 2021).

Regardless of the voluntary limitations on a global level or the ones imposed by regulators, the issue of hatred and disinformation will put the corporate business model to a significant test by new legal acts that are being prepared in the European Union, which are discussed in the next chapter. Additionally,

actions have been initiated in the UN system regarding online regulation that is modeled on the UN's global process of preventing climate change.

2 Fighting Disinformation: Civil Society

The system of power control must be based on truth, just as power itself must be based on truth and recognition of the values and indisputable things on which there should be a consensus. Disinformation campaigns strive to create a world of relativized values, chaotic reality, and a sense of uncertainty in people. This is why they target major democratic institutions such as elections, human rights, social and international solidarity, freedom of speech, and truth itself. Unlike with taxation and low-cost registration, there are no havens in this regard. For example, Malta was shaken by the murder of investigative journalist Daphne Caruana Galizia, who was tracking down corruption in the government and its connections to the business world. Everyone, including the Prime Minister of Malta, opposition representatives, and journalists involved in the investigation, reported hacking attacks and falsified emails, leading to information chaos and doubts about whether the truth behind her murder would ever be revealed (Malta: Journalists, 2021).

While some may argue that the potential for online violence is not high and that its impact is primarily individual (Van Dongen, 2021), the fact remains that the illegal annexation of Crimea and Russia's incursion into eastern Ukraine included a mass-scale disinformation component. By February 2022, the Russian-Ukrainian war had resulted in 14,000 deaths and tens of thousands of injuries. Once a full invasion began, at least twice as many were killed and injured in the first month alone. Therefore, while tragic events like the Capitol riots may be the result of independent, irresponsible actors, the impact of disinformation can have far-reaching and devastating consequences on a larger scale.

Disinformation is often propagated by media influencers who exploit human phobias for financial gain. Proposed measures must consider whether exposing, stigmatizing, and criticizing these individuals through journalistic investigations, social pressure, and the risk of losing credibility and profits may be effective deterrents. Similar methods used to combat hate speech, conspiracy theories, and online extremism should also be investigated to prevent foreign interference. However, several questions and dilemmas arise, including: How can we counteract disinformation without inadvertently strengthening it? How can individual researchers and journalists tackle this issue in a world where truth has become relative and trust in journalism is declining?

Civil society plays a critical role in combating disinformation, particularly within the research, education, and media communities. These groups not only assist in recognizing disinformation but also in understanding the nature of the problem. They offer expertise, counseling, and training to public service employees, but most importantly, they educate users and stakeholders in the information space. The scope of expert initiatives engaged in this work in Western countries is so vast that preparing a comprehensive map of these organizations and outlets could be the subject of a separate study.

In the international arena, American and British universities and research centers are highly influential due to their potential, resonance, scope, and impact. Many of these institutions collaborate with experts from other countries to combat disinformation. The American Atlantic Council, for example, has specialized teams that deal with disinformation, and their analytical work is used not only by the U.S. government but also by other countries and international organizations. The Center for European Policy Analysis, RAND Corporation, and the Brookings Institution also regularly provide analyses and recommendations for governments. The German Marshall Fund of the United States (GMF) has launched the *Alliance for Securing Democracy* project, which raises awareness of the dangers of disinformation, publicizes the results of scientific research, and regularly provides summaries of narratives and disinformation activities by Russia, China, and Iran on their website, *Hamilton 2.0 Dashboard*. The GMF also created a special project to analyze foreign narratives used during the election campaign in Germany in 2021.

Europe is not lagging when it comes to disinformation measures. The network of national and international organizations and institutions dedicated to fighting disinformation is constantly growing, and the cooperation between them is becoming increasingly fruitful. In Brussels, there are many specialized think tanks and initiatives focused on analysis, fact-checking, and education, including EDMO, EU DisinfoLab, and Lie Detectors. These organizations, together with institutions operating in EU member states, form networked communities that receive organizational and financial support from the European Union. Universities are also doing significant work individually or as part of international research clusters.

In Ukraine, *Stopfake* is one of the most effective networks for tracking and revealing disinformation. It is active in many countries and in many languages. It was established on the initiative of university staff and journalism students at the Mogilev Academy in Kiev (Stopfake, 2022). Many other activities were undertaken, first by the Ukrainians themselves and their organizations, then by others, including masses of ordinary internet users worldwide. The activities were aimed at combatting Russian propaganda and information falsehoods

related to the war in Ukraine. Some of these activities are unprecedented for their grassroots scale, as was the case with the group Anonymous massively attacking websites of Russian state institutions, including the Ministry of Defense and the secret services.

Selected examples of international activities undertaken in the field of countering disinformation by experts, researchers, journalists, and teachers, include:

- Bellingcat, the European Disinformation Media Observatory, the International Fact-Checking Network, all of which initiated and developed an international cooperation of researchers and journalists dealing with investigations and fact-checking.
- BBC, Lie Detectors, which focuses on education in schools.
- The Center for Countering Digital Hate, which unmasked and made a list of the 12 largest global COVID-19 related disinformers.
- The Center for International Resilience Detection, which detected a network of 350 accounts spreading powder propaganda in France.
- The German Marshall Fund of the U.S., which participated in pre-election monitoring focused on messaging by foreign actors.
- The Global Disinformation Index, which has diagnosed the most popular information portals in terms of disinformation potential in selected countries and researched the risk of disinformation and media credibility.
- GLOBSEC, which is developing a Decalogue of Transatlantic Principles for Combating Disinformation and is responsible for the regional initiative of the Alliance for Healthy Infosphere.
- Code for Africa, which is an international network dedicated to addressing technology, journalism, and fact-checking problems.
- DROG, Cambridge University, and the U.S. State and Homeland Security Departments, which facilitate international projects of training programs.
- Mandiant, which has identified perpetrators of hacking activities.
- MEMO 98, Who Targets Me, and Citizen D, which monitored election campaigns in Slovakia and Slovenia.
- The University of Oxford, University of Michigan, and Meedan, which researched and created of an algorithm facilitating the fight against disinformation on communication platforms.

The need for an integrated global approach to combating disinformation is demonstrated by the appeal that came from the milieu of European and American researchers. It aims to promote the document containing the *10 Transatlantic Principles for a Healthy Online Information Space* (*10 Transatlantic,* 2021):

1. Strive for greater transparency in the online information space.
2. Empower users to make informed decisions about their data.
3. Foster a culture of digital responsibility and accountability.
4. Minimize the spread of harmful information online.
5. Work towards timely, standardized, and proportionate rules for the digital space.
6. Support the ethical use of AI systems that embrace democracy and human rights.
7. Develop tools to increase citizens' media and digital literacy.
8. Empower civil society and the public to get involved.
9. Nurture an open space for competition to avoid monopolies.
10. Search for transatlantic solutions and beyond.

The call for international cooperation on the basis of these principles was initiated by the Slovak organization GLOBSEC. Its other regional initiative is the *Alliance for Healthy Infosphere,* which bring together think-tank centers from Central and Eastern Europe to combine their expertise and activities and counter disinformation more effectively.

The Propaganda Research Laboratory at the University of Texas at Austin, along with other experts, initiated the development of the *10 Transatlantic Principles for a Healthy Online Information Space*. Their two-year research focused on analyzing the network behavior of American propagandists working for political parties, national or foreign government agencies, or consulting or PR firms. The study found that manipulators use both coded platforms, such as WhatsApp and Telegram, as well as more open platforms like Facebook and YouTube, to influence minority voting behavior in specific states or cities. The groups targeted by manipulators include immigrant communities in swing states like Florida and North Carolina, where their voting behavior may sway presidential elections (Woolley & Sawiris, 2021).

Technological advancements in machine learning have made it possible to accelerate the tedious process of verifying information, such as fact-checking automation. Scientists from the University of Oxford have created a special algorithm that informs WhatsApp users if the message they received has been verified for authenticity (*Tackling misinformation*, 2021). This is an example of the closer cooperation between journalists, institutions, and organizations specialized in detecting disinformation, and the use of artificial intelligence. With increasingly advanced techniques and the cooperation of research centers, journalistic and expert circles are joining the fight against disinformation in various contexts, including elections, media education, tracking and exposing perpetrators, and studying the relationships

between disinformation and hackers. They are also creating maps of disinformation media and considering how to redesign the functioning of the internet and social media.

The following points synthesize important areas of activity from selected initiatives in the prevention and fight against disinformation.

2.1 *Elections*

Initiatives like MEMO 98, Who Targets Me in Slovakia, and Citizen D in Slovenia not only monitor election campaigns but also initiate legislative projects. Despite political forces defending themselves against transparency, their experiences and work show the scale of challenges regarding transparency and fairness of electoral processes. In many places in the West, parties, politicians, and the government's restraint leaves room for abuse in these matters. However, these challenges can be mitigated by international pre-election monitoring by global think-tanks. These include organizations focused on messaging by foreign authors appearing in the activities of broadcasters which may show further interference in the course of the campaign. The GMF *Alliance for Securing Democracy* project, mentioned above, contributed to observing the 2021 elections to the Bundestag while detecting and reducing foreign interference and disinformation.

2.2 *Media Education*

Research carried out by the RAND Corporation on countering disinformation has shown that most analysts propose changes in government policies, particularly to improve media education. While governments and supranational organizations like the EU can provide support, without increased participation from civil society, research, and journalistic communities, media education will likely remain inadequate in many countries. These communities possess valuable expertise and specialist skills that can be shared with educators. Initiatives such as Lie Detectors and Demagog.org, often created as a result of grassroots community efforts, have demonstrated effective ways to do this, including by incorporating games and play into mainstream education.

The potential of these lighter forms of fighting disinformation has been recognized not only by Cambridge University psychologists and Dutch activists, but also by the U.S. Departments of State and Homeland Security, who collaborated on the aforementioned game, *Harmony Square*. In the game, the player assumes the role of the "disinformation director" whose task is to sow discord and disrupt social harmony. The game, developed as part of an international project, uses the previously mentioned approach of psychological

inoculation. The game creators also developed similar products for children and teenagers.

2.3 *Attribution*

Bellingcat is known for conducting extensive investigations that go beyond exposing acts of disinformation. Their investigative potential, operating model, and ability to mobilize international cooperation make them a leading civil society institution. They have been instrumental in exposing the organizers of influence operations, including international disinformation campaigns like relating to the shooting down of Malaysian Airlines flight MH17 and attempts to poison Sergei and Julia Skripal. They have also uncovered manipulation of information during hostilities, such as those conducted by Russia in Ukraine. Other organizations, such as the Center for Countering Digital Hate, have compiled lists of the most influential and harmful global manipulators of information, including those spreading misinformation about the coronavirus. The Center for Disinformation Resilience (CIR) recently detected a network of 350 accounts spreading pro-China propaganda in the French-language information space (Carmichael, 2021), prompting YouTube and Facebook to take more decisive action against their infodemic activities.

2.4 *Mapping Disinformation*

The Global Disinformation Index (GDI) partnered with a Malaysian organization to identify the most popular portals in Malaysia with potential for disinformation by assessing the likelihood of encountering falsehoods (Media Market, 2021). This approach holds promise for detecting and preventing disinformation on a large scale. GDI has also conducted a media credibility study in other countries, such as Brazil, where half of the tested media were determined to have a high or very high risk of disinformation (*Disinformation Risk*, 2021).

Studies by the NATO Center of Excellence for Strategic Communication in Riga and the Political Capital Institute in Budapest provide a specific catalogue of media in Central and Eastern Europe, including Poland. These reports reveal an interesting pattern: media outlets with more "Balticness" in their names in Estonia, Lithuania, and Latvia and more "Polish", "national", or "independent" in their titles in Poland are more likely to have connections with disinformation activities, mainly by Russia (*Where to look*, 2017).

2.5 *Advanced Techniques*

States that organize disinformation campaigns often outsource their activities to third parties to conceal their involvement. However, investigative techniques and metadata analysis can reveal such connections and detect the

direct actors responsible for creating, for instance, a network of thousands of bot accounts or operating with an equally large number of accounts pretending to be real people within coordinated inauthentic behavior. Perpetrators also use programming techniques that confuse platform algorithms by simulating the authenticity of the account. Coordinated inauthentic behavior can be coupled with manipulation using genuine accounts and people who duplicate manipulated content, even in mainstream media. Identifying such activities requires not only knowledge of basic computer techniques but also advanced techniques, broader knowledge, and detection tools.

In disinformation operations, particularly the most dangerous ones, the key element may be hacking into email or social media accounts and then manipulating the publications of the stolen content. Methods of identifying the perpetrator of such hacking activities, including operation "Ghost Writer", which targeted Poland, were discovered by Mandiant, the analytical arm of the global cybersecurity company FireEye. Its reports indicated location certificates, specific use of Tactics, Techniques, and Procedures (TTPs), and phishing via e-mails when senders impersonate cybersecurity experts (Roncone et al., 2021).

2.6 *Cooperation*

International cooperation between researchers and journalists involved in fact-checking, pooling forces and resources, and initiatives like Bellingcat demonstrate that countering disinformation is not only time-consuming and labor-intensive but also costly. Networking is therefore an effective way to share resources and tasks. It can be assumed that joint efforts of researchers and journalists contributed to greater transparency in the presidential elections in France in 2017, the presidential elections in the United States in 2020, and the disclosure of disinformation, leading to the victory of a democrat in Moldova in 2021. For years, Germany was criticized for its passivity in responding to Russian disinformation, and in the parliamentary elections in 2021, many organizations in Germany finally cooperated to address the issue. While criticism certainly played a role, increased awareness of threats in the world of politics and government structures, intensified by pressure from civil society, also contributed to the actions taken.

Community cooperation, though most visible in Europe and North America, is beginning to extend to other continents. Code for Africa is the largest network of international cooperation focused on solving problems of media technology, journalism, and fact-checking (Knight, 2021). This is paramount in an interconnected and global world where continental, regional, and national weaknesses are eagerly exploited by disinformation organizers.

2.7 *A New Model of the Internet*

Researchers are increasingly focused on managing online platforms and their impact on people, regulatory needs, and the role of actors in this process, largely due to the problem of disinformation. Francis Fukuyama, an American scientist, has examined the idea of rebuilding and managing the internet in the modern era. He concluded that the power of platforms is so significant that they can determine the outcome of an election, and he advocated for reducing this potential. To achieve this, he proposed establishing middleware that would allow users to have greater control over the content they consume. Platform custodians, or librarians, would be at the center of this process, ensuring that information is based on knowledge and facts. Such structures should be integrated by the platforms themselves to avoid direct state interference (Fukuyama, 2021). This is an attempt to reconcile the freedom of the internet with consequences similar to regulatory effects. While implementing Fukuyama's ideas would likely lead to fragmentation in the network, it would give most users a chance to choose reliable sources more often than in an environment where the choice is left to the platforms through algorithms.

Samuel Woolley suggests that, to protect people from existing threats and problems, a new model of democratic internet management must be created. He believes that artificial intelligence is a challenge comparable to the control and non-proliferation of weapons of mass destruction (Woolley, 2021). The author proposes specific education for scientists, researchers, and data processing specialists to consider the potential effects of their implementation during the development of new digital solutions. Meanwhile, industry leaders like former Google CEO Eric Schmidt suggest in public statements that, similar to nuclear weapons, once the genie is released, it cannot be squeezed back into the same bottle. Even if the threat has passed its tipping point, however, the passivity of potential victims would only accelerate the emergence of further problems.

Since the mid-2010s, especially after the 2016 US elections, disinformation has garnered significant attention, resources, and cooperation. However, constructive critics argue that these efforts are often short-term and focused on pre-election periods. The actors involved in these efforts often do not work together or share the results of their work. Additionally, the mediatic nature of the topic means that lower-quality products can easily infiltrate the information space. Fragmentary research on the short-term effects of influence operations and overly simplistic attributions of responsibility undermine the effectiveness of countermeasures. To address these issues, there is a need for more funding, coordination, and development of verified research methods. This applies to Poland as well, where the Polish Institute of International

Affairs (PISM) and the Center for Eastern Studies (OSW) are among the centers with international reputations conducting and publishing regular research on disinformation in international relations.

3 Disinformation and Challenges for the Future of Journalism

Journalists are at the forefront of the fight against disinformation. The future of this profession depends on the credibility of the media and, more broadly, on how to make people willing to pay for reliable information. Traditional media is often seen as part of the establishment, making it difficult to reach younger audiences. The European Broadcasting Union emphasizes two key elements in countering disinformation: content moderation, which includes not only deleting posts but also addressing user complaints and ensuring transparency in their resolution, and law enforcement. While these are important steps, they represent a narrow approach that does not consider the nature of social media and its impact on the information environment. The enormity of the work undertaken by the media and journalists cannot be overstated, and they face significant challenges and threats. Research has shown that journalists covering COVID-19 were under the same level of stress as healthcare workers during the pandemic (Osmann, 2021). Despite this, their commitment to sharing the truth during the pandemic has led to a clear increase in trust in credible media in many countries, although unfortunately not on a global scale.

At the same time, fighting against disinformation requires determination and even boldness, which means providing help to those who dare to fight it. Jessikka Aro is one of the best-known examples of a journalist who felt alone in the face of pro-Kremlin aggression on the internet and decided to leave Finland. The journalist became a target of massive trolling, including threats from pro-Kremlin circles, against which she felt the Finnish authorities had provided insufficient protection.

Hate affects most journalists who deal with disinformation, and thankfully their work is highly appreciated, as in the cases of Maria Ressa and Dmitry Muratov, who were awarded Nobel Prizes. However, besides the journalism community, governments also have a primary responsibility to protect journalists' independence and security. Free media, truth-based science, and education are the cornerstones of democracy, and protecting them is crucial. The fight against disinformation starts with recognizing what it is. People and institutions, even those with greater authority, can make mistakes.

There are numerous examples of established media and state institutions responsible for public affairs disseminating misleading information. For

instance, the renowned German daily *Handelsblatt* repeated unverified information about the low effectiveness of the Oxford-Astra Zeneca vaccine among the elderly and erroneously stated that the U.S. epidemiological authorities had made premature decisions to cancel the obligation to wear protective masks.

In January 2022, when the tension around the Russian-Ukrainian conflict grew due to the expansion of Russia's military forces on the border with Ukraine and aggressive war rhetoric, Germany became the subject of mass criticism for what it did not do. They did not refuse to fly British planes with military equipment for Ukraine over their territory as they did not receive requests for permission.

Journalists often face the challenge of avoiding the spread of disinformation while maintaining their credibility. One approach is to listen to the concerns of the audience and prioritize transparency and authenticity in their reporting. Artificial intelligence can also play a role in improving journalism, but it must be used appropriately. Journalists must work to engage with audiences who may feel excluded and avoid reinforcing disinformation while reporting on it. It is widely agreed that simply negating false information without providing context or explanation can inadvertently spread it further. This is a tactic used by disinformers and some politicians who prioritize exposure over accuracy. The guiding principle for journalists therefore be "first of all, do no harm".

There is a growing sense of the need to revise certain complex concepts such as journalistic neutrality in light of the prevalence of lies and fake news. During the tenure of former U.S. President Donald Trump, one of the most prolific producers of fake news in recent years, the old truth that what the head of state says is news had to be revisited. For example, American journalistic dilemmas surfaced prominently regarding attempts to amend the electoral law to be called publicly electoral restrictions, as the Democrats wanted, or fairness laws, as the Republicans wished (Hobes, 2021). In other words, comparing the incomparable and attempting to forcefully look for arguments to balance criticism and offer the rostrum to liars is not an effective approach. Journalists need to consider how not to fall into the trap of symmetries. BBC journalist Rebecca Skippage (2021), head of the monitoring team, evoking the thoughts of Hannah Arendt, directly referred to such a war of narrative: "Talk in an engaging way, like the bad guys". This is one of the main messages for journalists who wish to fight false information.

There are three basic premises and tasks from the point of view of journalistic effectiveness in counteracting disinformation.

The first premise is responsibility and trust. The media and journalists remain at the forefront of the fight against disinformation. For many of them, there are important, often personal, issues at stake, including survival in the

profession. The state should support them in these efforts, finance media education with their participation, guarantee reliable public and private media independence, and maximize access to recipients. The main challenges posed by misinformation and the temptation to engage in unethical journalistic practices are concentrated within the media industry, acting as a lens to magnify these issues. The controversial self-promotion tactics employed, for instance, by the UK's *GB News* network, while technically permissible, undermine public trust in mainstream media more broadly (Orr, 2021). Additionally, pro-Kremlin trolls exploit the fundamental right to freedom of expression in the West by leaving comments on articles published in major Western news outlets. These comments are then used as propaganda in Russia to support the Kremlin's disinformation campaigns. This is a new method of Russian-controlled disinformation that has escalated in recent years as social media platforms have intensified their efforts to combat it.

It is therefore essential to have reliable content moderation on the internet, not only for social media platforms but also for traditional media outlets. Moreover, traditional media should be cautious when downgrading their content standards on social media to increase their reach at the cost of their message's reliability.

The issue of credibility and trust in traditional media is also linked to how some social groups, particularly younger generations, perceive them as an "us versus them" scenario. The "us" represents the rebellious, alienated, and disadvantaged, while the "them" represents the wealthy elite. To build trust, journalists must engage directly with these groups by participating in media education programs in schools and local communities.

The second premise is fact-checking, which involves complex activities that require cooperation and time. Dozens of media outlets and hundreds of journalists have collaborated on some international journalistic investigations. From an efficiency and credibility standpoint, therefore, it is particularly important to join forces, pool resources, and share information. The trends in media development or the crisis in traditional media will make fact-checking one of the key trends in journalism. Its main goal is to improve the quality of public debate, which is contrary to what information manipulators want. This includes verifying the statements of politicians, officials, or other influential people who find their way into the public sphere. Fact-checking is necessary, but it is not sufficient on its own. Its corrective function is implemented *post-factum* when the damage has already occurred. Research has shown that the human mind is often resistant to late corrections, and exposing disinformation has varying degrees of effectiveness depending on the recipients.

Fact-checking can help to reduce the impact of manipulated narratives and disinformation on people's attitudes and beliefs. However, it may not always change their support for a particular political party. In such cases, pre-empting disinformation by providing civic education and encouraging people to identify unreliable or biased sources may be more effective. Renowned research centers like the NATO COE in Riga or the Global Disinformation Index can assist in analyzing and mapping these problems.

To date, fact-checking has not been profitable, which has limited the resources available for its development. However, stakeholders such as Google and Facebook do pay for outsourcing, with Facebook paying around $2 million in 2019. Fact-checking services like Full Fact in the UK or Liberation in France receive around $200,000 per year from these stakeholders.

In recent years, mainstream media has invested more resources into fact-checking. Agence France Press's (AFP) Fact Check, which started as a one-person unit in 2017, has grown into a team of 120 people working in 24 languages and 80 countries as of 2021. The practice of fact-checking is evolving into a more digital and efficient process. In fact, fact-checking played a significant role in the context of the 2021 parliamentary elections in Germany. Funded by a Google grant, DPA and FaktenCheck21 trained 600 journalists from 100 media organizations. Despite the challenges facing this field of journalism, the public demands its growth and development (Scire, 2021)

Despite what some people claim, fact-checking does not have to be counterproductive (Grabmeier, 2021). In fact, by denying untruth, fact-checking, if done properly, can help to dispel falsehoods without amplifying them. That is why disinformers in Spain have attempted to discredit journalists and fact-checking itself by impersonating fact-checkers in the eyes of confused recipients (Loucaides, 2021). Crowdsourced fact-checking, which involves engaging ordinary internet users to verify information, has also shown its strength as an innovative method. However, there is a risk that this activity, while filled with good intentions, may not necessarily involve professionals with the necessary skills.

The third premise is media education. As previously mentioned, media education is a neglected subject in schools in the vast majority of countries, including many in Europe, due to the lack of curricula and teaching competencies. To address this, the media community and journalists can be involved in training educators or delivering the programs themselves. Funding for such activities could come from a portion of tax revenues, such as an audio-visual media tax, and combined local and regional government resources. In the UK, the BBC's *School Report* and in Finland, journalists' *Faktana, kiitos!* programs

involve a growing number of students. Similar initiatives in France, such as *Entre les lines* in partnership with *Le Monde* and AFP, have also gained traction. International efforts to pool resources have also been successful, such as LieDetectors, a Brussels-based non-governmental organization that conducts media education programs in schools across many EU countries, including Germany and France.