Israel Public Policy Institute

HEINRICH BÖLL STIFTUNG TEL AVIV

Embassy of the Federal Republic of Germany Tel Aviv

Policy Paper

Disrupted Democracies

A Multistakeholder Approach to Fight Disinformation and Online Political Manipulation

Rafael Goldzweig

Facing up to the Infodemic: Promoting a Fact-Based Public Discourse in Times of Crisis

Policy Paper Series by the Israel Public Policy Institute (IPPI)

Disrupted Democracies

A Multistakeholder Approach to Fight Disinformation and Online Political Manipulation

Author

Rafael Goldzweig

About this Paper

This policy paper is part of the paper series "Facing up to the Infodemic: Promoting a Fact-Based Public Discourse in Times of Crisis."

Against the backdrop of the COVID-19 crisis, this paper series explores some of the key challenges facing democratic societies as a result of misinformation in the digital public sphere. It features a unique mosaic of perspectives and insights by experts from Israel and Germany that shed light on different facets of the phenomenon of online misinformation, with the aim of invigorating a societal debate on the issue as well as offering concrete ideas about how to address it.

The series "Facing up to the Infodemic: Promoting a Fact-Based Public Discourse in Times of Crisis" was generously supported by the German Embassy in Tel Aviv. The content and opinions expressed in the papers are solely of the authors and do not necessarily reflect the views of the German government and/or the Israel Public Policy Institute.

About the Project

This paper series is part of the broader project "Fostering Democratic Resilience in the Digital Age," conceptualized and executed by the Israel Public Policy Institute (IPPI) in collaboration with the Heinrich Böll Foundation, Tel Aviv.

The objective of the project is to promote dialogue, exchange of knowledge and collaboration between researchers and practitioners from Israel and abroad to enhance democratic resilience in the context of the changing media and information landscape in the digital age.

Please cite as follows:

Goldzweig, R. (2021). Disrupted Democracies: A Multistakeholder Approach to Fight Disinformation and Online Political Manipulation. Policy Paper Series by the Israel Public Policy Institute: "Facing up to the Infodemic: Promoting a Fact-Based Public Discourse in Times of Crisis."

Contents

Executive Summary

This policy paper departs from the straightforward observation that the public debate in countries across the globe is gradually shifting from the offline "town square" to the online sphere, where polarization and false information often shape (and radicalize) political views. Manipulation using disinformation as a strategy might serve political and financial interests of actors bent on influencing citizens' perceptions in key moments such as elections and national political discussions, effectively disrupting democratic systems, both domestically and abroad.

Against this backdrop, this paper explores the information gap between the tech companies, regulators and civil society, arguing that the online platforms alone cannot provide the answer to the complex phenomenon of online manipulation. The paper draws on questions of regulation with a focus on the European legislative debate, emphasizing the need to promote a multi-stakeholder approach as a way to increase societal resilience to disinformation and effectively consolidate democratic foundations to withstand emerging online threats in the future.

As a way forward, the paper makes the following recommendations:

- **Policymakers** ought to provide regulatory frameworks to increase online platforms' obligations and accountability when it comes to transparency about actions taken to curb online disinformation and hate speech. Regulation should treat disinformation as a societal problem and consequently go beyond regulating social media platforms, i.e. by introducing additional measures such as strengthening media pluralism, empowering citizens and funding civil society.
- **Civil society organizations** must also play a significant role, for example by actively monitoring what happens on online platforms, identifying manipulation attempts, and countering hybrid threats such as disinformation and hate speech in real time.
- **Tech companies** should improve the way they enforce community standards, increase transparency of content moderation and ranking/recommender algorithms and share data with civil society organizations that work to identify manipulation attempts.

1. Introduction

The use of personal data in an unregulated environment over the past 20 years has given rise to new business models that use consumer data to predict and shape behavior. Data use has unlocked markets and is at the core of most of the competitive companies, defining business and marketing strategies across different industries. However, while such models have optimized companies' targeting strategies, and to some extent benefited consumers through offering them more tailor-made services and products, they have introduced possibilities for abuse. The same mechanisms used to target consumers – segmentation, and micro-targeting – have been used to target voters during important elections, shedding light on the challenges that new technologies pose to democracies.

One of these new markets enabled the success of social media platforms – online spaces created in the early 2000s where users connect to each other, consume information, and share personal content. Their monetization strategy is based on the idea that the more a company knows about you, the more it can predict users' choices and behavior, and use this information to influence consumer practices. To this end, algorithms craftily select content to make users stay as long as possible in platforms that collect their data; by spending time on these platforms, the user is also exposed to the ads that keep this machine running.

The same business model that made these companies successful provided a ready-to-use infrastructure to intensify polarization through disinformation and hate speech shared at scale, ultimately influencing political behavior and arguably playing a role in democratic processes. This tactic was introduced into the mainstream in 2016, placing social media platforms at the center of a controversy about their role in allowing malicious actors to use their services for political manipulation.

This manipulation was made possible by the increasing number of users of these services in the past years. The COVID-19 pandemic accelerated an already observed trend: tech companies are becoming central to public discourse, as citizens are increasingly using such platforms to vent their feelings and share their political ideas, including during the course of elections. With offline discussions paused in many countries for the time being, such exchanges are happening in online spaces subject to manipulation.

The COVID-19 pandemic accelerated an already observed trend: tech companies are becoming central to public discourse, as citizens are increasingly using such platforms to vent their feelings and share their political ideas, including during the course of elections.

In this environment, disinformation has developed as an unsavory means to push for political and financial interests. Russia's use of digital tools to influence other countries' internal politics is now old news: China, Iran, and Saudi Arabia have also established themselves in this field. Recently, the European Union accused China of being behind a huge wave of COVID-19 disinformation campaigns aimed at weakening how governments respond to the pandemic.[1] According to the Oxford Internet Institute, evidence of organized social media manipulation campaigns was observed in 70 countries in 2019, up from 48 countries in 2018 and 28 countries in 2017.[2]

While it is clear that public opinion manipulation using social media platforms was performed by state and non-state actors, measuring the effects of the exposure to false information and hate speech online continues to be a challenge. Psychology researchers point out several reasons why social media and the current way in which we consume information makes us prone to believe in narratives pushed by disinformation campaigns.[3] These include the use of mental shortcuts and tendency to consume information that reinforces one's pre-existing beliefs.[4] However, measuring whether and to what extent pieces of false information online have the power to alter a person's decision on how to vote is a difficult task.

This policy paper explores the challenges posed by this gradual shift of the public sphere from the "town square" to online spaces that are designed and moderated by private companies. It will explore how the new gatekeepers – namely big tech companies such as Facebook, Google, Twitter and others – whose solutions are focused on just certain manifestations of the new tactics of attention manipulation, are not positioned to provide all the answers to this problem. The added effect of polarization and disinformation cannot be solved unless there is a society-wide response.

Pointing towards a framework to increase societal resilience against such hybrid threats, the article explores possible regulatory responses to address this issue, focusing on the European debate, and the desired role of civil society in this new digital public sphere.

2. The Online Public Sphere

The public sphere is the domain of social life where public opinion can be formed (Habermas,1991), namely, the space where individuals come together to discuss and identify societal problems, influencing social mobilization and political action.[5] With the advent of Web 2.0, platforms connecting and giving voice to citizens were seen as the ultimate level of democracy, empowering those who were not able to overcome the barriers imposed by traditional gatekeepers – media, elites and political figures.

From the early days of MySpace to the advent of TikTok, the role of social media has changed. MySpace, in 2004, was the first platform to reach the number of one million active users.[6] Now, in 2020, an estimated 3.6 billion people are using social media around the world, a number projected to increase to almost 4.41 billion in 2025.[7] The majority of such users exchange messages and engage in political debates over platforms owned by three companies: Tencent (WeChat), Facebook & Google, with Bytedance (TikTok) growing fast, and other platforms operating in specific niches and contexts (Twitter, Telegram, Reddit, among others). Online exchanges underwent a transformation from text to link, and are now rapidly evolving into image and short video exchanges.

MySpace, in 2004, was the first platform to reach the number of one million active users. Now, in 2020, an estimated 3.6 billion people are using social media around the world, a number projected to increase to almost 4.41 billion in 2025.

Throughout this period, connectivity facilitated collective action, where online calls for action and events turned into offline protests and revolutions, leading to actual political change. Viral messages on Twitter and Facebook sparked the Arab Spring across different autocracies in 2011, leading to democratic change in countries such as Tunisia, or deepening conflicts as in Libya. Activists' clever use of information and communication technologies (ICTs) for organizing and sustaining the Euromaidan protests in Ukraine in 2014 was seen as a key element that allowed the revolution to keep going, ultimately resulting in the removal of pro-Russia President Yanukovych and the call for new elections.[8] For better or for worse, governments and actors around the world understood the power of a connected society, of a public sphere where anyone could voice their uncensored opinions and mobilize for political action. Gaining momentum online could lead to offline change.

It is worth noting that the fast pace at which digital tools evolved in shaping our habits in the past decades interplayed with pre-existing political developments of different societies. In the United States, for example, political polarization has been increasing steadily since the 1960s,[9] with studies pointing out to increasing societal polarization in demographic groups that do not use social media platforms in large numbers.[10] In other parts of the world, the growth of ethnic, political and religious conflicts did not stop over the past decades,[11] with some studies pointing to benefits of connectivity to affected areas.[12]

The centrality of social media platforms put them in the spotlight. While not being the root cause of trends that predate the creation of such platforms, the rapid privatization of the public sphere became an opportunity for state and non-state actors to instrumentalize their (geo)political interests by manipulating online services. This trend of abuse has been accelerating steadily in different parts of the world, with an increase in the number of countries engaging in social media manipulation from 28 to 70 in the last three years.[13]

The rapid privatization of the public sphere became an opportunity for state and non-state actors to instrumentalize their (geo)political interests by manipulating online services.

The above-mentioned cases show that platforms are different in the role they play in this digital public sphere, but one thing that they have in common is that their design shapes how people exchange information. The different designs offer different possibilities for manipulation.[14] These platforms also lack transparency, scrutiny, and boundaries for what is allowed or not, which opens them up to abuse by malicious actors. With the world moving ever faster to the online environment in the aftermath of the COVID-19 pandemic, the process of digitizing the public sphere accelerated, and counting on platforms alone to provide the answers proved to be insufficient.

3. Can Tech Companies Alone Provide the Answers?

The Brexit referendum and US elections in 2016 brought to the world's attention to how social media platforms could be weaponized for (geo) political purposes. Whether they were used to attack specific political opponents or to spew false information at scale, social media manipulation ended up eroding trust in facts and democratic values that were said to have been strengthened by giving people the voice they had lacked a few years earlier. Cambridge Analytica shed light on how easy it was to use personal data to shape political perceptions. It showed that targeting mechanisms aimed at influencing consumers can be used in the same way when it comes to influencing voters' political choices.

New forms of online manipulation and hate followed. In 2017, in a milestone case of unsupervised abuse of the online space, the Independent International Fact-Finding Mission in Myanmar stated that "Facebook has been a useful instrument for those seeking to spread hate" against the Rohingya, a Muslim minority in the country.[15] The report recognized that Facebook had tried to respond to misuses of its platform, but found it "slow and ineffective." Large scale ethnic violence against Rohingya Muslims in Myanmar dates to 1978,[16] but the abuse of new technologies helped to give a new face to an old problem. At the time that the new wave of violence broke, Facebook had 7.3 million active users in the country, but only around four Burmese speakers reviewing content.[17]

While 2018 was a year when the "tech giants" dealt with several instances of abuse around the world, it was also the year when encrypted platforms started showing their potential for massive manipulation. WhatsApp got the spotlight for its use as a channel for massive disinformation sharing during the Brazilian presidential elections,[18] an attempt to skew voters' perceptions ahead of the vote. The encrypted characteristic of this platform was exploited to send false information at scale to different groups, which in turn became viral with people sharing with their own contacts.

While 2018 was a year when the "tech giants" dealt with several instances of abuse around the world, it was also the year when encrypted platforms started showing their potential for massive manipulation.

Another messaging app, Telegram, made headlines for its role in Belarus, due to its being used to spark protests after the fraudulent landslide re-election of Alexander Lukashenko in August 2020. The app allows for users to follow channels, which pushed the so-called "March of Freedom," a series of opposition protests in the aftermath of the elections.[19] The platform became the gateway to organizing protests in a scenario of political instability.

This short overview paints a picture of the dual role of social media: they connect and empower users, reduce barriers for political participation and empower voters, but they also show how vulnerable we are, albeit in different ways, to massive manipulation by state and non-state actors. In this new digital version of citizenship, we have moved from offline venues for discourse to online environments that allow us to connect to fellow citizens and mobilize change around us. At the same time, those very environments are exposing us to false information, hatred and conspiracy theories, which can lead to further polarization in society and extremism in politics.

Online environments allow us to connect to fellow citizens and mobilize change around us, but at the same time, are exposing us to false information, hatred and conspiracy theories, which can lead to further polarization in society and extremism in politics.

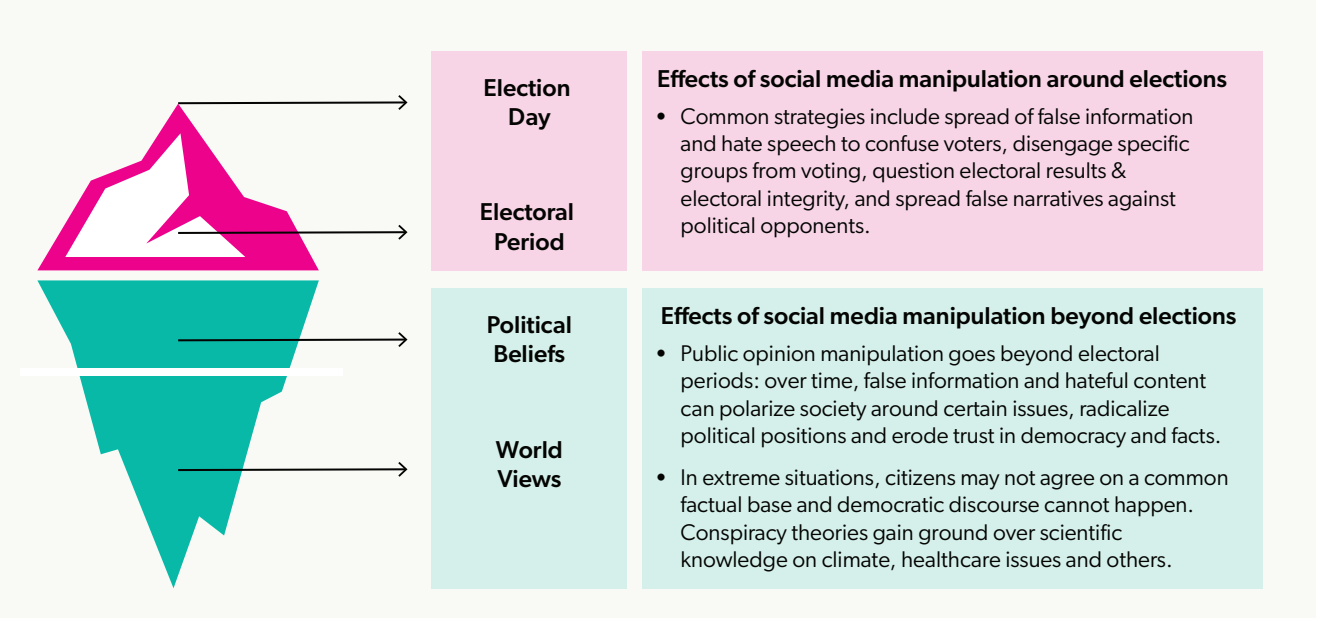
The different examples show that among the many effects of social media, there is one particular growing concern about the effect of such platforms on society, namely, the role of such platforms during big political moments, such as elections, revolutions or crises. However, the cases above illustrate that the exploitation of social media for political ends and its effects extend far beyond those "big moments" and have also medium and long-term impact on users' political behavior and beliefs.

The major tech companies have invested massively in the past years to preserve the integrity of elections and prevent their platforms from becoming tools for political manipulation and the spread of harmful content. New products to identify and limit the spread of false information were created, economic incentives disrupted, data shared with researchers and authoritative information made more easily accessible on major social media platforms.[20]

The major tech companies have invested massively in the past years to preserve the integrity of elections and prevent their platforms from becoming tools for political manipulation and the spread of harmful content.

Figure 1.

Intended Effects of Social Media Manipulations Around and Beyond Elections



However, the fact that more false accounts have been taken down and malicious activity has been clamped down is not a sign that matters are improving; the widespread attempts and varied techniques still used in political manipulation also mean that more actors are becoming active in trying to manipulate such platforms. Moreover, there has been growing evidence that the effects of social media manipulation go far beyond elections, playing a role in the long-term effects on political opinion formation.

The rising coordination and strength of extremist groups and conspiracy theories on different social media platforms have shown that radical voices are finding ways to use the online public sphere to convince other citizens to believe in a different set of facts. Facebook recently removed 790 groups linked to QAnon conspiracy theories,[21] but on other platforms such as Telegram or Parler, several QAnon channels and groups are free to operate and spread their narratives without much oversight. More recently, QAnon conspiracy theories have been reinforced by believers of COVID-19 falsehoods and climate change deniers.[22]

> **Some companies are taking stronger actions than others; they have different rules as to what is allowed, and different capacities to enforce these rules.**

This presents a complex challenge. Some companies are taking stronger actions than others; they have different rules as to what is allowed, and different capacities to enforce these rules. Also, polarization and radicalization happen around the clock, requiring extensive monitoring that could become costly, even if a company has the desire and capacity to carry it out. Lastly, placing the ultimate responsibility of monitoring public debates online to a few private companies raises issues of legitimacy: should they be the main actors responsible for this task, or should it be a more collective responsibility?

Indeed, outsourcing the solution to problems affecting public trust in politics and democracy brings several issues into relief. The Russian influence on the 2016 elections, and the case of Myanmar, where Facebook did not provide proper oversight,[23] are just a few examples demonstrating that tech companies cannot fight this complex problem alone. Four years later, both the 2020 US elections and the 2020 Myanmar elections showed a much more coordinated and organized response from tech companies than in 2016. Without civil society involvement including the momentum it generated in the media, however, these efforts would have not been successful.

In 2020, a global disinformation challenge joined the mix. The COVID-19 outbreak confronted all countries with the "infodemic" – false information about the virus, ranging from non-existent cures to falsehoods about how countries were dealing with the pandemic. The platforms' attempts to limit the spread of the infodemic suggest that tech companies can weed out misinformation – provided that a source of authoritative information is clearly defined.[24] In the case of the COVID-19 crisis, platforms relied on scientific evidence provided by the World Health Organization.[25]

> **The platforms' attempts to limit the spread of the infodemic suggest that tech companies can weed out misinformation – provided that a source of authoritative information is clearly defined.**

As a result of experience gained during the pandemic, many platforms have tightened their rules on advertising, prohibiting ads that create a sense of urgency in the context of the coronavirus, such as ads implying a limited supply of medical gear, or advertising substances that supposedly cure the infection. Facebook announced it would notify users who had engaged with misinformation about the virus. Together with Twitter and YouTube, the company took down content that could cause harm, including posts by Brazilian President Jair Bolsonaro and Venezuela's Nicolas Maduro that praised a dubious cure for COVID-19 and encouraged ending social distancing measures. WhatsApp has limited message forwarding options on its platform in an attempt to reduce the spread of misinformation about the virus. Twitter has started labelling tweets that contain deceptive or manipulated content, while Google has directed searches on the virus to reliable websites and taken down Google Play apps promising information about the pandemic that was not approved by a national government or medical institution.

The lessons learned from previous elections, as well as the COVID-19 infodemic, is that online platforms play a central role in shaping public discourse and providing users with relevant information about a range of public discussions, from elections to a global pandemic. The platforms' responses to such events also show that different companies invest different efforts in preventing these spaces from being manipulated. At the end of the day, while they can indeed do more to prevent their services from being abused, constantly increasing their activity, companies cannot solve pre-existing conflicts and issues in countries they operate in, or be expected to predict all malicious or unintentional uses of their services. Making them the ultimate responsible actors to protect democracy from interference and abuse of their services for (geo)political gains is a simplification of a complex problem.

What is necessary is a multi-faceted solution involving a whole set of actors. In recent years, policymakers, civil society organizations, journalists, fact-checkers and international electoral observer groups became vocal in this debate, each offering a different angle to the problem, without offering a single effective solution. Developing and developed democracies alike faced decreased societal trust in facts and democratic institutions and observed the rise of extreme voices in their mainstream political systems. The next sections will consider approaches to a solution: How can policymakers and civil society play a role in this complex scenario?

4. Framing the Regulation Debate

Policymakers have often used regulation to address emerging challenges, and their response to those posed by the use of social media for political purposes have proved no different.[26] In different parts of the world, several regulatory models have been suggested that propose outsourcing the responsibility for dealing with problematic content to the companies that facilitate this content, or empowering courts to decide on the legality of online content, and even criminalizing users that produce or spread this content.

The current relationship between regulators and industry is characterized by information asymmetry. Tech companies are currently best positioned to offer solutions because they have more effective means and knowledge to deliver faster responses to rapidly evolving challenges; governments lack the necessary expertise. However, given that the manipulation of social media platforms poses grave dangers to democracy, it is imperative to involve non-industry actors in this pursuit of solutions, by demanding greater transparency on the part of tech giants. Disinformation is not a

technological problem, but a societal one, and as such, the asymmetry in knowledge between tech companies and civil society is perilous to democracy in the long run.

Disinformation is not a technological problem, but a societal one, and as such, the asymmetry in knowledge between tech companies and civil society is perilous to democracy in the long run.

Legislators, regulators and civil society organizations need to increase their technical understanding in order to find effective remedies, and create ecosystems where decisions of tech industries are scrutinized and discussed. However, this framework should not interfere with the capacity and resources of companies to generate quick responses to disinformation and hate speech transmitted over their platforms. Hence, striking a balance is an essential and complex task for legislators.

Regarding the regulation of tech companies, two main areas are currently being explored by legislators. The first relates to rules governing tax revenues, data protection, market dominance and privacy standards. Advances have been made in legislation dealing with these questions: espionage scandals and several breaches of users' data helped increase awareness and several countries approved data protection laws, most prominently the GDPR in the European Union. Other discussions have been conducted on the aspect of antitrust law, to discuss whether big tech mergers constitute a form of market dominance.[27] The second area, which is the focus of this section, applies to content shared on such platforms and the abuse of such platforms for political interests that can be harmful for democracy in the long run. Regulation dealing with content includes rules and procedures regarding hate speech, political polarization, the role of political advertising and mis/disinformation online. While different models and approaches have been tested, none so far have proved to be a silver bullet for dealing with the problem, and in some parts of the world, faulty regulation has been used as a weapon against freedom of expression, with the clear intent of shutting down opposing voices in society.[28]

In the context of elections, social media companies partnered with local independent fact-checkers to have a third party responsible for determining the quality of the content before they took any action. Platforms are right when they say that they shouldn't be the ones deciding what content should be permitted in the first place; while terrorism and child pornography may be easier topics to address through legislation, disinformation and propaganda may be more subtle and evasive.

While laws defining what content should or should not be allowed may not be desirable, this does not mean that there are no rules. Social media platforms have codes of conduct called community standards, or terms of service, which guide rules and behaviors accepted by the platform and aim to set boundaries of what is and is not permitted in these virtual communities. We can think of them as self-regulatory actions to prevent abuses. They have been changing over the years to respond to the abuse of platforms around elections, or to the sharing of problematic content via their platforms. To enforce these codes, platforms use a mix of AI identification methods and human reviewers to decide what content to allow and what to take down.

Different companies have varying definitions for these problems, or adhere to a varying levels of transparency about how their services are abused, leading to inconsistencies in the rules applied to different platforms.

What has been happening in recent years is that some of these community standards have not been appropriately enforced, and malicious actors have created new techniques to bypass or exploit weaknesses in platform standards. Also, different companies have varying definitions for these problems, or adhere to a varying levels of transparency about how their services are abused, leading to inconsistencies in the rules applied to different platforms.

To account for the apparent lack of enforcement by tech companies, some countries have tried to tighten the standards to be followed by such companies. The Macron administration introduced

Box 1.

Political Advertising in Different Tech Platforms

- **Google** defines election ads as ads that feature a political party, a current elected officeholder, or candidate, or that pose a referendum question, a referendum campaign group, or a call to vote related to a national referendum or a state or provincial referendum on sovereignty.
- **Twitter** prohibits the promotion of political content, defining it as content that references a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome. It does not, however, mention whether it allows paid content dealing with social issues that could be used to polarize public opinion without necessarily referencing these concepts.
- **Facebook** has a broader concept, defining them as "Ads About Social Issues, Elections or Politics," where social issues go beyond the traditional electoral angle, covering sensitive topics that are debated and may influence public opinion in other moments.

This multitude of definitions creates different standards when dealing with this type of content,[29] and transparency requirements are subject to the discretion of private companies. The inconsistency is reflected, for example, in ad transparency requirements:

- **Google** provides information on ads purchased on elections in the EU & UK, India, New Zealand and the US.[30]
- **Twitter** decided to block political advertising altogether around the world, but continued allowing issue ads in all countries but the US.[31]
- **Facebook** offers by far the most complete information about political and issue ads, with the Ad Library[32] feature covering almost all countries and territories, and the Ad Library Report[33] covering 71 countries at the time this study was conducted.

In this context, legislation can be helpful for developing a baseline definition of what constitutes political advertising, as well as baseline criteria for transparency.

a law outsourcing the control over the spread of misinformation during elections to the courts, which failed to pass the Senate in 2018. A similar failed approach in Italy tried to impose fines on users who shared false information online. The most notable example, however, comes from Germany's Network Enforcement Act, in effect since 2017, which holds social media platforms accountable for the content they facilitate, demanding illegal hate speech content to be taken down according to the country's criminal code.

> **Legislators around the world are now faced with the following question: how to create effective regulation that increases transparency and accountability across different platforms, while preserving the fundamental rights of freedom of expression.**

The fragmentation of concepts and standards across different platforms, together with the different levels of transparency and enforcement required by different countries, leaves the path for exploitation wide open across the globe. Legislators around the world are now faced with the following question: how to create effective regulation that increases transparency and accountability across different platforms, while preserving the fundamental rights of freedom of expression. A further challenge pertains to fragmentation: how can regulators ensure compatibility of standards internationally so that tech companies and new market entrants can operate across different countries, while at the same time empowering civil society and media to address systemic causes that allow disinformation to flourish? Answering these questions independently and without coordination could result in each state dealing with such issues differently, depending on its legislators' understanding of the problem. This could result in a collection of divergent rules that require companies to operate differently in each country, but do little to solve the problem as a whole. The EU discussions on the subject therefore offer some avenues for considering a regional approach to these challenges.

4.1. European Debate: The Digital Services Act and the European Democracy Action Plan

At the EU level, regulation has not kept pace with the technological changes of the past few decades. The 2000 e-Commerce Directive established a general regulatory framework for tech companies and defined, among others, transparency and information requirements and voluntary codes of conduct for online service providers.[34] However, the purposes and types of digital service providers have changed drastically in the 20 years since the directive was adopted. The EU has since issued numerous acts of fragmented regulation, policies and codes of conduct on issues related to governing the internet and the digital sphere, but a general updated framework regulating the current challenges technology presents to democracy is sorely lacking. Concomitantly, member states have adopted their own regulations to address these rising challenges, creating a situation of regulatory fragmentation across the EU that runs counter to a coordinated and cohesive approach.

Yet, to address the rising challenges posed by online disinformation and hate speech, European countries took a co-regulatory approach, based on the possibility of adopting voluntary codes established by the E-Commerce Directive. In May

2016, the EU Code of Conduct on Countering Illegal Hate Speech Online[35] was adopted to prevent and counter the spread of illegal hate speech online. Later, in an attempt to counter potential influence campaigns ahead of the 2019 European Parliamentary elections, the EU Commission adopted the Code of Practice on Disinformation.[36] Both codes established transparency standards to be followed by companies, requiring them to report on how they were handling disinformation and hate speech, and what actions they were taking to fight them through the enforcement of their community standards.

It is worth mentioning that as the mechanisms established under the Code of Practice on Disinformation were designed with the European Parliamentary elections in mind, they neglected to provide tools addressing the broader, systemic effect of disinformation campaigns, which exceeds "big political moments." False information influences users of different age groups on different platforms on which it is shared, affecting not only voter behavior, but political beliefs and world views (see Figure 1, above).

> **The Codes of Practice and Conduct adopted at the EU level established voluntary monitoring practices that proved a good first step, but they were insufficient for fighting the broader effects of hate speech/illegal content and disinformation.**

The Codes of Practice and Conduct adopted at the EU level established voluntary monitoring practices that proved a good first step, but they were insufficient for fighting the broader effects of hate speech/illegal content and disinformation.

They lacked enforcement mechanisms to address whether tech companies are meeting their responsibilities, and the narrow focus of these frameworks did not result in an ongoing and systematic effort to prevent platforms from being manipulated by foreign or domestic actors with deleterious political or financial agendas.

Aiming at closing this gap and building on the experience and lessons learned,[37] the Commission suggested coordinating standards to ensure smooth digital operations across the common market, by updating the e-Commerce directive through the Digital Services Act (DSA).[38] The Act – currently a legislative proposal that the European Parliament will vote on in 2021 – will be reinforced by a complementary set of non-legislative measures included in the European Democracy Action Plan (EDAP),[39] which will focus on democratic resilience, electoral integrity and a comprehensive set of actions to combat disinformation.

The DSA aims to deepen the internal market for digital services, by increasing and harmonizing the responsibilities of online platforms and by bolstering the oversight over platforms' content policies. Aside from aiming to ensure uniform, fair and contestable markets for new entrants, this regulation package also intends to spare social media platforms and other relevant companies from a fragmented legal environment, especially as some member states (most notably France and Germany) have already adopted national legislation on illegal content and hate speech. One of the key areas the DSA intends to address is increased transparency in the operation of tech platforms. Under the DSA, transparency requirements will likely cover two elements: benchmarks, and verification mechanisms.

Regarding benchmarks, the lessons learned

from the Codes of Practice and Conduct will likely inform the definition of specific indicators for measuring and perfecting transparency standards. These should include data about the type of content that is restricted, blocked, or removed, as well as continuous supply of information to regulatory agencies and civil society on how much content has been taken down, on what grounds and within what timeframe, the number of appeals, their resolutions, and other details.

Verification mechanisms would include the involvement of third parties to increase external accountability, including through mandating companies to share data with researchers. The European Data Protection Supervisor's preliminary opinion on data protection and scientific research states that EU codes of conduct for scientific research could be used to push for access to data that enables enforcement of data-protection standards while addressing the needs of researchers, who have been requesting data from tech companies and were, in many instances, denied access to it. In addition to third-party verification, financial and other support of data journalism and oversight organizations who research tech companies' actions would help to develop a specialized public that could scrutinize companies' reporting.

Third-party verification assumes the existence of a functioning and active civil society, and much of what will ensure its existence is contained in the European Democracy Action Plan. In contrast to the single-market logic of the DSA, the EDAP is framed within the context of defending European democracies, specifically from the manipulation of public opinion and disinformation campaigns aimed at undermining democratic institutions. Its focus is the protection of the EU against foreign interference and the large scale spread of disinformation and hate speech on digital platforms, especially around elections. The EDAP also calls for greater transparency of information, highlights the importance of trust in the media in the EU, and in addition to its other instruments, includes a specific action plan targeting the media and audio-visual sector.

> **Regulation alone is insufficient and undesirable for solving the complex challenges posed by new technologies; it needs to be followed by a series of non-legislative measures to increase societal resilience.**

The two initiatives have different goals, but they will complement and reinforce one other. Their interaction demonstrates an interesting architecture to approach challenges posed by public-discourse manipulation: regulation alone (DSA) is insufficient and undesirable for solving the complex challenges posed by new technologies; it needs to be followed by a series of non-legislative measures (EDAP) to increase societal resilience.

The EDAP package[40] aims to reinforce three integrated themes, for which citizen engagement will be of particular importance:

- Ensure that electoral systems are free and fair and preserve electoral integrity;
- Strengthen freedom of expression and the democratic debate through media freedom and media pluralism, and an active civil society;
- Tackle disinformation in a coherent manner and build upon actions listed in the recent communication on tackling COVID-19-related disinformation.[41]

Such a framework, an initiative of the European Commission, understands disinformation and defense of democracy in the digital age as a shared responsibility, and appreciates the need for balance in ensuring media pluralism while enabling civil society to play a more active role in this environment. Around the world, governments have been trying to come up with a silver bullet to tackle disinformation and online challenges to democracy through legislation alone, only to find that the solution might be more dangerous than the problem itself. Such an approach presents an alternative by providing conditions for civil society to play a stronger role, as well as providing conditions for media pluralism to be strengthened.

5. Role of Civil Society and Research Institutions

The challenges posed by new technologies lie in **how** malicious actors might use these platforms for their own interests. They might be financially motivated, buying or creating artificial engagement to push for a brand, or monetizing a website through a click-baiting strategy. Alternatively, and often more problematic, they can be politically motivated. As mentioned before, past examples include foreign interference in important political moments of another country, such as elections, through hate speech shared at scale against minority groups or extreme groups aiming at polarizing public debate in their own countries.

The fact that interference in public debate online can come from domestic actors, as well as international ones, renders the role of civil society and research institutions even more important, especially given the international focus of EU strategizing to date. The EU StratCom Task Force, an arm of the European External Action Service, was initially tasked to counter Russia's ongoing disinformation campaign, and recently, in light of COVID-19-related disinformation, it turned its focus to China. In other words, the EU mandate is to look at actors trying to manipulate public opinion coming from outside of the block. Neither the EU nor its member states would find acceptance in acting as monitors of public discourse online in their own countries. After all, government interference in the online debates could be seen as state surveillance or censorship, and as of now, the different standards of the rule of law within member states would enable prosecution on charges of anti-democratic interference if these actions were to be fulfilled by states.

> **Civil society plays a vital strategic role – it has different mandates and is not bound by government interests, freeing it to provide a transparent assessment of how false information is used to manipulate public opinion.**

Therefore, civil society plays a vital strategic role – it has different mandates and is not bound by government interests, freeing it to provide a transparent assessment of how false information is used to manipulate public opinion. For that, it needs sustained funding and access to data for research in order to understand the different ways in which disinformation and hate speech are generated and disseminated. Another advantage of civil society is that it can view the issue through different angles, taking, for instance, a gender-harassment perspective[42] or monitoring in greater detail the different aspects of online interference during elections,[43] such as whether it is motivated by foreign interests or coming from interest groups inside of a given country.

This role can also be played by investigative journalists, fact-checkers with digital forensics skills and independent researchers. Since online manipulation and hybrid threats are new strategies, countering them opens different fields of work to be filled and developed by different actors. To this end, civil society and researchers need sustained funding to update their scales, but also to perform investigations and advocacy. Only an informed and active civil society is capable of shedding light on these new strategies and effectively countering them.

> **Partnering with organizations that look at the issue from different angles, such as universities, think tanks, fact-checkers and civil society organizations, is an effective way to shed light on manipulation campaigns and advocate for specific changes to transparency standards or legislation.**

Four years after the unexpected interference in the 2016 US elections, the Election Integrity Partnership[44] – a group composed of the DFRLab, Graphika, Stanford Internet Observatory and University of Washington – was created to serve as a knowledge hub regarding the forces shaping public opinion online, and as a watchdog in anticipation of future manipulation attempts ahead of the 2020 Elections. Such responses can be easily replicated, and to some extent have been observed in different elections since 2016. Combining efforts and partnering with organizations that look at the issue from different angles, such as universities, think tanks, fact-checkers and civil society organizations, is an effective way to shed light on manipulation campaigns and advocate for specific changes to transparency standards or legislation.

6. Policy Recommendations and Next Steps

This policy paper has discussed the complexity of dealing with disinformation and related hazards of the digital age. **Social media platforms** have a global reach and have been investing in solutions, but the effort to counter threats posed by online manipulation must be a shared one. After pointing out the areas in which social media companies have fallen short over the past years, this policy paper explored the role of legislators and civil society in ensuring that democracies are protected in the near and long term.

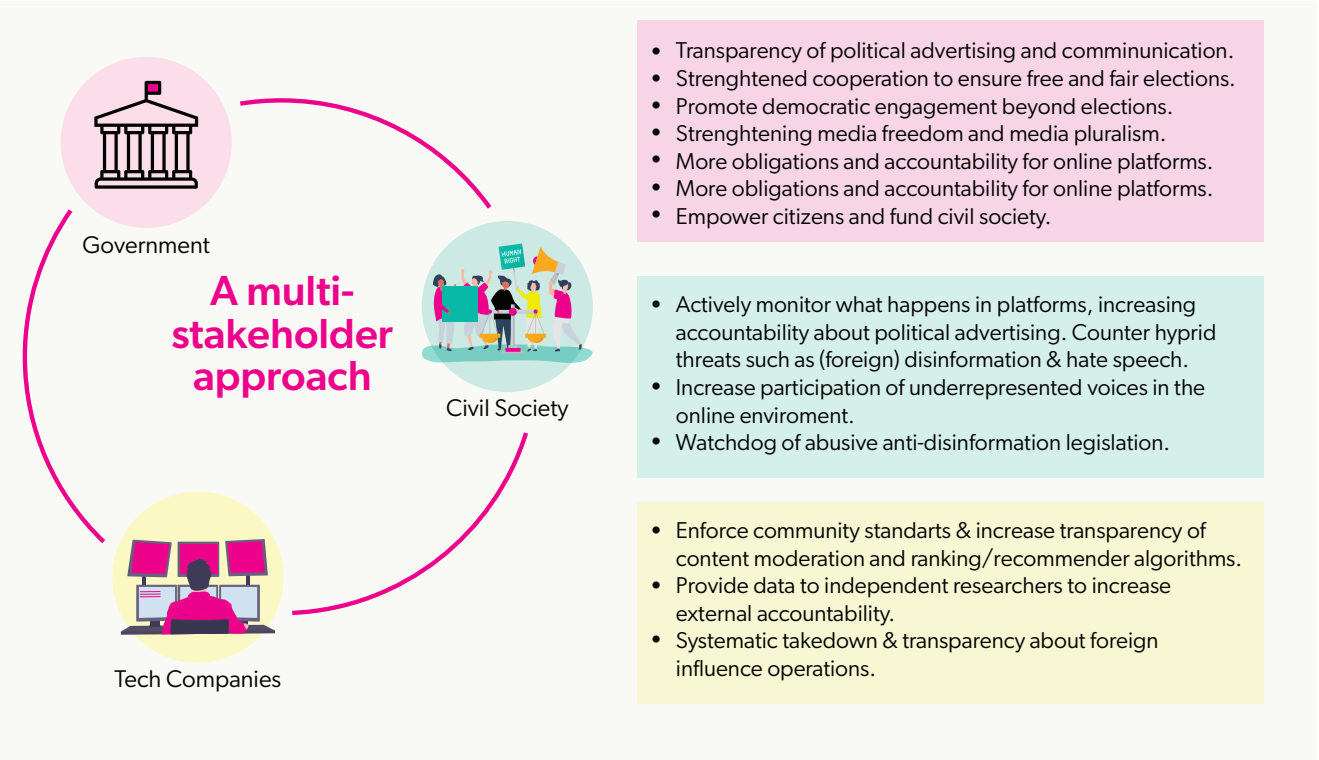
The **role of the state** is central, not only in standardizing operative rules for different tech platforms, but also in demanding transparency requirements and creating an ecosystem that will act as a system of checks and balances to ensure damage control. It is not desirable for governments or courts to take a pivotal role in defining what is allowed and not when it comes to the grey zones posed by disinformation. In countries where the rule of law fails to properly control the abuses of the forces in power, such legal mechanisms can be counterproductive, as they can be abused to attack journalists or activists.

> **Civil society, researchers, and fact-checkers need to have a framework in which they can cooperate and where their voices are heard.**

For all of the above, **civil society, researchers, and fact-checkers** need to have a framework in which they can cooperate and where their voices are heard. A pluralistic civil society that is well funded and outfitted with activists who possess the necessary skills and data to perform their tasks is a central element to ensure that malicious

Figure 2.

Recommendations for Lawmakers, Civil Society and Tech Companies



actors are not successful in their attempt to misinform users, preventing them from radicalizing public opinion in the long term.

It is worth noting that the mix of solutions is broad, and should be thought in the short, medium- and long-term perspectives. One angle of intervention is bolstering media literacy. In the long term, media literacy has the potential to educate a new generation of users who are more tech savvy and aware of the potential for deception that exists in the online space. Literacy will have to confront the continuing evolution of technology, given the likelihood that future innovations might pose different threats than those we are currently used to – notably the use of artificial intelligence to create content, such as video (deepfakes), images, text and voice creation.

> **A multi-pronged model for ensuring integrity of online public discourse, during and beyond elections, one that is to some extent replicable in different countries, is imperative.**

In conclusion, a multi-pronged model for ensuring integrity of online public discourse, during and beyond elections, one that is to some extent replicable in different countries, is imperative – not only to counter hybrid threats, but in order to understand how they evolve and change across different regions of the world, so to better protect against them in the future. The existing information gap between tech companies, civil society and governments causes

undue and perilous delays in the urgent task of developing an appropriate response to rising challenges posed by technology. A common societal understanding, as a result of data sharing between with researchers and greater transparency about how malicious actors abuse such services, will help build resilience to the manipulation of public opinion as we move deeper into the online public sphere.

The conflicts of the 21st century will inherently have an online aspect, and such spaces might be weaponized to threaten national sovereignty, democracy, conflict resolution and human rights. A multi-stakeholder approach is the only way to ensure that solutions are balanced and proportional to the risks.

References

[1] Rankin, J. (2020). "EU says China behind 'huge wave' of Covid-19 disinformation," The Guardian, https://www.theguardian.com/world/2020/jun/10/eu-says-china-behind-huge-wave-covid-19-disinformation-campaign

[2] Bradshaw, S. & Howard, P. N. (2019). The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation. Computational Propaganda Research Project, University of Oxford. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf

[3] Shane, T. (2020). "How Does our Psychology Make Us More Vulnerable to Misinformation?" First Draft News, https://firstdraftnews.org/latest/the-psychology-of-misinformation-why-were-vulnerable/

[4] Lorenz-Spreen, P. (2021). Human Cognition and Online Behavior During the First Social Media Pandemic: Breaking Down the Psychology of Online Information Consumption in the Context of the COVID-19 Pandemic. Policy Paper Series by the Israel Public Policy Institute: "Facing up to the Infodemic: Promoting a Fact-Based Public Discourse in Times of Crisis."

[5] Habermas, J. (1991). The structural transformation of the public sphere: An inquiry into a category of bourgeois society. Cambridge, Mass: MIT Press.

[6] Ortiz-Ospina, E. (2019). "The Rise of Social Media." Our World in Data. https://ourworldindata.org/rise-of-social-media

[7] Clement, J. (2020). Number of social network users worldwide from 2017 to 2025, Statista. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/#:~:text=Social%20media%20usage%20is%20one,almost%204.41%20billion%20in%202025.&text=Social%20network%20–penetration%20is%20constantly,2020%20stood%20at%2049%20percent.

[8] Bohdanova, T. (2014). "Unexpected Revolution: The Role of Social Media in Ukraine's Euromaidan Uprising, European View," Sage Journals. https://journals.sagepub.com/doi/full/10.1007/s12290-014-0296-4

[8] Lewis, J. (2020). "Polarization in Congress," Department of Political Science (UCLA). https://voteview.com/articles/party_polarization

[10] Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). "Is the Internet Causing Political Polarization? Evidence from Demographics." Brown University. https://www.brown.edu/Research/Shapiro/pdfs/age-polars.pdf

[11] Dupuy, K. & Rustad, S. A. (2018). "Trends in Armed Conflict, 1946–2017," Peace Research Institute Oslo (PRIO). Conflict Trends, 05/2018. https://reliefweb.int/sites/reliefweb.int/files/resources/Dupuy%2C%20Rustad-%20Trends%20in%20Armed%20Conflict%2C%201946%E2%80%932017%2C%20Conflict%20Trends%205-2018.pdf

[12] Hiller, P. (2018). "How Social Media is Changing Conflict." Peace Science Digest, https://peacesciencedigest.org/social-media-changing-conflict/

[13] Bradshaw & Howard (2019).

[14] Lorenz-Spreen (2021).

[15] UNHRC (2017). Independent International Fact-Finding Mission on Myanmar. https://www.ohchr.org/en/hrbodies/hrc/myanmarffm/pages/index.aspx

[16] Hussain, M. (2016). Rohingya Refugees Seek to Return Home to Myanmar. Voa News. https://www.voanews.com/east-asia-pacific/rohingya-refugees-seek-return-home-myanmar

[17] Stecklow, S. (2018). Inside Facebook's Myanmar operation Hatebook. Reuters. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

[18] Roose, M. & Isaac, K. (2018). Disinformation Spreads on WhatsApp Ahead of Brazilian Election. NYTimes. https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html

[19] Williams, S. (2020). Belarus has torn up the protest rulebook. Everyone should listen. Wired. https://www.wired.co.uk/article/belarus-protests-telegram

[20] Facebook (2020). Working to Stop Misinformation and False News. https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news; Twitter (2020). Election Integrity Policy. https://help.twitter.com/en/rules-and-policies/election-integrity-policy; TikTok (2020). Combating Misinformation and election interference on TikTok. https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok; YouTube (2020). How YouTube supports elections. https://blog.youtube/news-and-events/how-youtube-supports-elections.

[21] Frenkel, S. (2020). Facebook Removes 790 QAnon Groups to Fight Conspiracy Theory. NYTimes. https://www.nytimes.com/2020/08/19/technology/facebook-qanon-groups-takedown.html

[22] Spring, M. & Wendling, M. (2020). How Covid-19 myths are merging with the QAnon conspiracy theory. BBC, https://www.bbc.com/news/blogs-trending-53997203

[23] Stevenson, A. (2018). Facebook Admits It Was Used to Incite Violence in Myanmar. NYTimes. https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

[24] Goldzweig, R. (2020). It is time tech companies act on election-time disinformation. Al Jazeera. https://www.aljazeera.com/indepth/opinion/time-tech-companies-act-election-time-disinformation-200520135809708.html

[25] World Health Organization (2020). WHO partners with WhatsApp, Facebook and Viber to bring most up to date and accurate information to billions of people. https://www.who.int/news-room/feature-stories/detail/who-partners-with-whatsapp-facebook-and-viber-to-bring-most-up-to-date-and-accurate-information-to-billions-of-people

[26] Goldzweig, R., Wachinger, M., Stockmann, D., & Roemmele, A. (2018). Beyond Regulation: Approaching the challenges of the new media environment. Dahrendorf Forum. https://www.dahrendorf-forum.eu/wp-content/uploads/2018/12/Beyond-Regulation_Final.pdf

[27] EDRi (2020). Competition law: Big Tech mergers, a dominance tool. https://edri.org/our-work/competition-law-big-tech-mergers-a-dominance-tool/

[28] In Russia, legislation was adopted to restrict access to information containing falsehoods, insults and disrespectful messages related to the symbols of the Russian Federation, including its authorities. Library of Congress (2020). Government Responses to Disinformation on Social Media Platforms: Russia. https://www.loc.gov/law/help/social-media-disinformation/russia.php

[29] For a comprehensive overview on how to effectively regulate political advertising, check Jaursch, J. (2020). Rules for Fair Digital Campaigning. Stiftung Neue Verantwortung. https://www.stiftung-nv.de/en/publication/rules-fair-digital-campaigning

[30] Google Transparency Report (2020). Political advertising on Google. https://transparencyreport.google.com/political-ads/home

[31] Twitter Transparency Center (2020). https://ads.twitter.com/transparency

[32] Facebook Ad Library (2020). https://www.facebook.com/business/help/2405092116183307?id=288762101909005

[33] Facebook Ad Library Report (2020). https://www.facebook.com/ads/library/report/ (accessed 9 September 2020)

[34] EU Commission (2020). e-Commerce Directive. https://ec.europa.eu/digital-single-market/en/e-commerce-directive

[35] EU Commission (2016). The EU Code of conduct on countering illegal hate speech online. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[36] EU Commission (2018). Code of Practice on Disinformation. https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[37] EU Commission (2020). Study for the assessment of the implementation of the Code of Practice on Disinformation. https://ec.europa.eu/digital-single-market/en/news/study-assessment-implementation-code-practice-disinformation

[38] EU Commission (2020). The Digital Services Act package. https://ec.europa.eu/digital-single-market/en/digital-services-act-package

[39] EU Commission (n.d.) European Democracy Action Plan. https://ec.europa.eu/info/strategy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan_en (accessed 9 September 2020).

[40] European Parliament (2020). Legislative Train Schedule. A new push for European democracy. https://www.europarl.europa.eu/legislative-train/theme-a-new-push-for-european-democracy/file-european-democracy-action-plan

[41] European Commission (2020). Joint communication to the European parliament, the European council, the council, the European economic and social committee and the committee of the regions. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020JC0008&from=EN

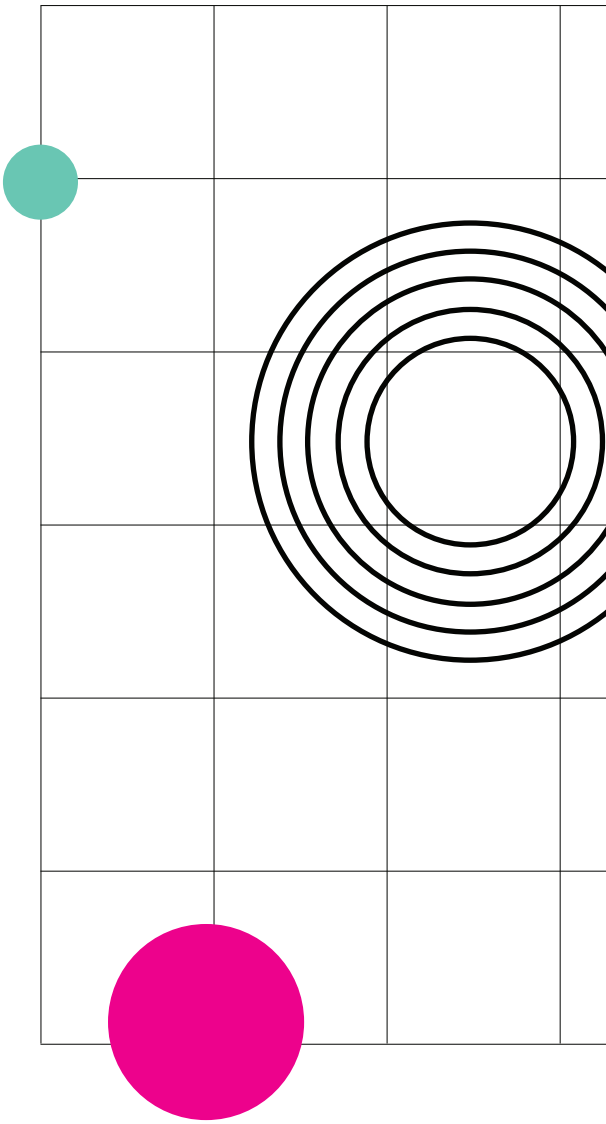
[42] Democracy Reporting International (2020). Gender & Social Media: Measuring Underrepresentation & Harassment. https://democracy-reporting.org/dri_publications/gender-social-media-measuring-underrepresentation-harassment/

[43] Supporting Democracy (2019). Guide for Civil Society on Monitoring Social Media during Elections. https://democracy-reporting.org/wp-content/uploads/2019/10/social-media-DEF.pdf

[44] Election Integrity Partnership (2020), https://www.eipartnership.net/

About the Author

Rafael Goldzweig researches the intersection of new technologies and elections, having worked on projects analyzing the impact of social media on elections in Tunisia, Sri Lanka, Libya, Myanmar, Brazil, Ukraine and several EU countries. He co-authored the "Guide for Civil Society on Monitoring Social Media During Elections," as part of the European Union funded project "Supporting Democracy." Previously, he worked and researched the intersection between new technologies and electoral behavior for the German NGO Democracy Reporting International, the Israel Public Policy Institute (IPPI), the think tank Dahrendorf Forum, and as a Google Policy Fellow in Panama City. He holds a Bachelor degree in International Relations (University of São Paulo) and a Master of Public Policy (Hertie School of Governance).



Project Partners: Fostering Democratic Resilience in the Digital Age

The paper series is published as part of the broader project "Fostering Democratic Resilience in the Digital Age", conceptualized and executed by the Israel Public Policy Institute (IPPI) in collaboration with the Heinrich Böll Foundation Tel Aviv.

Israel Public Policy Institute

Israel Public Policy Institute (IPPI)

The Israel Public Policy Institute (IPPI) is an independent policy think-and-do-tank and a multi-stakeholder dialog platform at the intersection of society, technology and the environment. Through its research activities, knowledge sharing, networking and public outreach, IPPI contributes to the innovation of public policy with the goal of understanding, guiding, and advancing the transformation process of our societies towards a sustainable and democratic future. IPPI works with a global network of actors from government, academia, civil society, and the private sector to foster international and interdisciplinary cross-pollination of ideas and experiences.

HEINRICH BÖLL STIFTUNG TEL AVIV

Heinrich Böll Foundation Tel Aviv

The Heinrich Böll Foundation is an independent global think-and-do-tank for green visions. With its international network of 33 international offices, the foundation works with well over 100 project partners in more than 60 countries. The foundation's work in Israel focuses on fostering democracy, promoting environmental sustainability, advancing gender equality, and promoting dialog and exchange of knowledge between public policy experts and institutions from Israel and Germany.

גרמני ישראלי DIALOG

German-Israeli Dialog Program of the Heinrich Böll Foundation

The German-Israeli Dialog Program of the Heinrich Böll Foundation was established to foster cooperation and exchange of knowledge between public policy communities from Germany and Israel with the aim of generating new actionable insights in support of democratic values and sustainable development. The program is home for a range of projects and activities that provide unique collaborative spaces for researchers and practitioners from government, academia, tech and civil society to meet, debate and formulate innovative policy-oriented solutions to societal questions and challenges shared by both countries.

Israel Public Policy Institute

→ office.israel@ippi.org.il
→ www.ippi.org.il

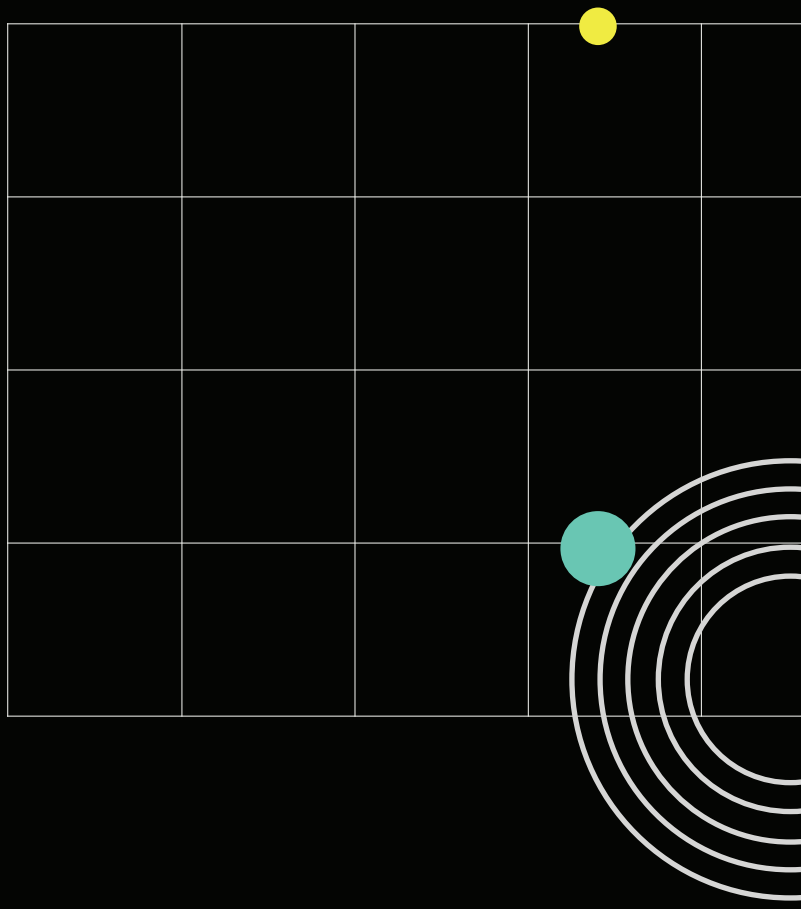
Heinrich Böll Foundation Tel-Aviv

→ info@il.boell.org
→ www.il.boell.org

German Embassy Tel Aviv

→ pr-s1@tela.diplo.de
→ tel-aviv.diplo.de/il-de

Release date: January 2021



Israel
Public Policy
Institute

HEINRICH BÖLL STIFTUNG
TEL AVIV

Embassy
of the Federal Republic of Germany
Tel Aviv