Journal of Political Science and International Relations
2024, Vol. 7, No. 4, pp. 75-88
https://doi.org/10.11648/j.jpsir.20240704.11

Research Article

A Sociopolitical Approach to Disinformation and AI: Concerns, Responses and Challenges

Pascaline Gaborit[*]

Project AI4Debunk, Pilot4dev, Brussels, Belgium

Abstract

International organizations classify disinformation as one of the main threats to democracy and institutions for more than a decade. Digital technologies reinvent and profoundly transform modern lifestyles, citizens' and business environments. AI is bringing a new disruption in the way we access knowledge and create, spread and understand information. It can also blur the lines between real information and manipulated information with the emergence of 'Fake News', automatic networks' cross referencing, and 'Deep Fakes'. AI systems enhance the potential for creating realistic fake content and targeted disinformation campaigns. Disinformation goes beyond simple rumors to deliberately deceive and distort evidence-based information through fabricated data. European institutions have also recently focused on the identification of disinformation linked to FIMI: Foreign Information Manipulation and Interference. The article identifies trends and concerns related to disinformation and AI. It explores the perception of disinformation, its impacts, and responses including the EU AI Act and online Platforms' policies. It provides a first analytical approach to the topic based on the current debates by researchers, the first findings of our 2024 surveys, interviews and the analysis of hundreds of online fake news items. It attempts to understand how citizens and stakeholders perceive disinformation and identifies possible impacts. It also analyzes the current challenges and constraints, opportunities and limitations to tackle manipulation and interference. The article considers the current processes, and impacts of disinformation (2), the presentation of the main findings of our online survey on the perceptions of disinformation (3), the current EU regulatory responses (4) and the Discussion Points (5). We argue in this article that there is a gigantic change in the way that we access information, but that the responses to disinformation are still at an early stage. The article also demonstrates that there is an increased awareness in European countries about the impacts of disinformation, but also a gap between the ability to identify "fake news" and disinformation, and a limited understanding of the processes, threats, and actors involved in spreading disinformation.

Keywords

Disinformation, Fake News, Misinformation, Artificial Intelligence, Manipulation, Interference, Social Media

1. Introduction

International organizations classify disinformation as one of the main threats to modern lifestyle and democracy for more than a decade. Digital technologies relentlessly reinvent and profoundly reshape modern lifestyles and business environments, and AI is bringing a new disruption into the way to access knowledge and to create, spread and to understand the

[*]Corresponding author: Pascaline.Gaborit@yahoo.com (Pascaline Gaborit)

Received: 2 October 2024; **Accepted:** 17 October 2024; **Published:** 31 October 2024

information by blurring the lines between real information and manipulated information. This digital revolution is reaching its golden age, in continuity with previous transformations which occurred within less than a decade. By 2024 the number of world mobile phone owners is forecasted to reach 7, 21 billion. Around 67% of the world population has currently access to the Internet and it was only 1% in 1995. In the next stage of technology, the creation of online collaborative platforms allied to social media, the 4G mobile phones, smart devices, and Internet cloud, have generated additional transformations (connected devices, instant connections, human-device interactions, new forms of information and images, the creation of online communities). Changes have loomed in quickly, without the possibility for people to step backwards, think them over, neither for all stakeholders to adapt skills, education courses nor to convert manifold economic sectors.

The advantages for consumers and business investors of the digital upheaval are colossal: direct access to worldwide information, knowledge and data is made possible…Easy-to digest knowledge is accessible almost anywhere through social media, internet search engines, selective applications but also online education. Another important transformation is that social media represent an important source of news for most of their users [43]. Connectivity is boundless: collaborative platforms including social media have enabled a direct link among people, but also between potential businesses entrepreneurs and consumers. Logistics is substantially facilitated by possibilities to order supplies, to move faster or cheaper, to use GPS, geo-localization and instant connections.

This brings us to the downsides and stumbling blocks, if not threats of this digital revolution on the possible manipulation of information. Indeed, the rapid evolving technologies, including with AI, are "increasing opportunities to create realistic AI-generated fake content, but also, (…) facilitating the dissemination of disinformation, to a (micro) targeted audience and at scale by malicious stakeholders" [8]. Concerns have been raised on copyrights, biased algorithms, business models using massive data to deceive individuals and replacement of jobs and employment by technology in numerous AI sectors. AI technologies will also facilitate the use of video, text, and image generating content based on false information, and creating difficulties for individuals and for the media to trust the information [50]. The disinformation can take different forms, including fake news but also impersonation enabled by deep fakes. is then relayed by bots and amplifiers through automatic dynamic cross-referencing of networks [5, 3, 47].

The project AI4debunk, led by the European Union aims at creating tools and methodologies to 'debunk disinformation, misinformation and fake news' in the social media, with the use of artificial intelligence. The project enabled the identification of more than 1000 case studies extracted from fact checkers as well as identified by the media partners of the project. The two case studies that were approached by the project were: disinformation on the war in Ukraine for the first one and on Climate change for the second case study. In addition to this, the project enabled the study of existing literature on the following topic: definitions, threat actors, polarizing narratives, and threads of disinformation. The first elements of the project are published on the project's website: www.ai4debunk.eu. Finally, the project enabled our research team to organize interviews with stakeholders, stakeholders' focus groups and the online survey which results are presented in point III.

The aim of this paper is to provide a first analytical approach to the topic based on the current debates by researchers and media-literature and media articles- and on a first analysis of fake news. The analysis will continue and be narrowed throughout the lifetime of the project – four years.

This paper does not address technologies as such, but rather brings some elements from political and social science towards an analytical framework which will help to understand what is at stake, before we are able to develop the different systems of 'debunking'. It will go through the current Processes, threats, and impacts of disinformation (2), the presentation of the main results of our online survey on the perception of disinformation (3) the current regulatory responses from the European Union (4) and the discussion points (5). We argue in his article that we face a gigantic change in the way that we access information, but that the responses to disinformation are still at an early stage. The article also demonstrates that there is an increased awareness in European countries about the impacts of disinformation, but also a discrepancy between the identification of 'fake news' and disinformation, and the lack of understanding of the processes, impacts, and actors of disinformation.

2. Processes, threats and Impacts of Disinformation

There have been questions on the trust and credibility in online activities since a decade [34]. The study on disinformation has been amplified since 2016 with the scandal created by the interference of Cambridge analytical in the U.S. elections. The topic has gained increased attention, especially in what is often referred to as the "post-truth era," a term that has gained popularity following the 2016 U.S. presidential election, during which "fake news" became a common phrase [35]. 'Fake news' is here understood as a general term used by the media to embrace both disinformation and misinformation. Identifying the processes, evidence of disinformation, and impacts is important for the understanding of the research.

Disinformation is defined as the deliberate dissemination of false, incorrect or misleading information to cause harm. Disinformation is false, inaccurate or misleading information that is shared with the intent to deceive the recipient [1, 8, 10, 51, 57, 59]. Misinformation can be defined as the deliberate dissemination of false, incorrect or misleading information which is not intentional. In a more simplistic way, the most important distinction between information and misinformation and dis-information would be the question of truth. Where information is true, misinformation or disinformation are untrue [10, 62]. Disinformation is not a new phenomenon; it has deep historical roots, stretching back to ancient times when rulers and leaders would intentionally spread rumors or misleading information to weaken opponents or control public opinion [27, 6, 59]. While modern technologies have dramatically increased the speed and scale at which disinformation can be spread, the tactics remain strikingly similar relying on manipulating emotions and sowing confusion to achieve a specific agenda. Before delving into the various policies and efforts aimed at regulating misinformation and disinformation, it is essential to first identify the processes, actors, narratives and impacts.

The European Commission defines disinformation as “verifiably false or misleading information that is created, presented, and disseminated for economic gain or to intentionally deceive the public, and may cause public harm. Public harm comprises threats to democratic political and policy-making processes as well as public goods such as the protection of EU citizens’ health, the environment or security. Disinformation does not include reporting errors, satire and parody, or clearly identified partisan news and commentary.” [21] By disseminating disinformation online, malicious stakeholders may for instance seek to discredit scientists or leaders, to polarize information or to destabilize democratic institutions. The indicators of disinformation are intentional harm; false or misleading content; presence of bias or manipulative techniques; audience targeting including micro targeting and psychometric profiling. Disinformation can take different forms, which is not restricted to the social media although social media manipulation is an important aspect of it. Apart from the social media manipulation it also encompasses false flag operations, influence operations and the manipulation of social media.

The Manipulation of social media, in particular, aims to create and amplify false or misleading narratives in messages spread through social media platforms. Among other examples, trolls or even trolls’ farms using social bots have been implicated in spreading divisive content on the main social media platforms like Facebook and X (former Twitter) and Telegram, aiming to sow discord and influence public opinion [5, 41, 55]. Although this manipulation content is not always traced back to its origin, experts have identified several ‘campaigns’ of manipulation on social media, some of which can be traced back to outside of the EU. The role of the Kremlin for instance in creating disinformation campaigns about the war in Ukraine, has uncovered several Telegram channels in Russian amplifying and spreading misleading messages [14]. The disinformation campaigns have more broadly been targeting to stir anger and reactions on divisive elements such as the war in Ukraine, the arrival of migrants, or on the conflicts in the Middle East.

Although most of the platforms have created moderation, this has not prevented the spread of false information, with a higher role of TikTok, X and Telegram because of their policies to restrict moderation to a minimum. However, Meta, Instagram and even traditional media have not been spared by the campaigns.

Interestingly, most of the moderation is now automatized, and according to Meta (Facebook, WhatsApp), more than 90% of the moderation is done by AI tools, the rest being done by human moderation [36]. The moderation did however not prevent massive disinformation campaigns, including media spoofing, impersonation of celebrities [5] and deep fakes, notably during the Doppelganger disinformation campaign which started in 2023 [55].

Malicious actors have engaged in false flag operations, where they pose as individuals or groups from different countries to spread disinformation and conceal their true identity. This tactic aims to exploit existing tensions and manipulate perceptions of geopolitical events. These false flag operations are expected to be empowered with the use of AI [5, 66]. The role of Large Language Models and automated dynamic network cross referencing are already massively used by the actors of disinformation [5, 41]. Worryingly, further advances in machine learning will increasingly enable adversaries to identify individuals' unique characteristics, beliefs, needs, and vulnerabilities. This will allow them to deliver highly personalized content and target those most susceptible to influence with maximum effectiveness [36]. These techniques could also create micro approaches, to target decision makers or voters. Disinformation has been making use of different narratives and undermines Western institutions and democracies. This has taken the form of fake news, but also taken the form of false/forged journal and media covers, use of voices, manipulation of images, and use of artificial intelligence. An example is the forgery of the satiric French Journal ‘Charlie Hebdo’ in February 2024 to mock the Ukrainian army command. This occurred simultaneously with the forging of covers of the newspapers Titanic and El Jueves[1]. This trend has enabled the emergence of fake news detectors which have been set up rather as a ‘reactive’ approach as they cannot ‘prevent’ the disinformation to be spread. Another trend is also the increasing acceptation of the use of ‘fake news’ and disinformation as a ‘political weapon’ among democracies as we unfortunately witness in European and in U.S. elections campaigns in 2024.

Several Countries have employed influence operations to shape perceptions and policies in target countries through a combination of disinformation, propaganda, and covert activities [11, 12]. These operations often target vulnerable populations and exploit societal divisions to advance their interests. Entities from foreign countries have been suspected of interfering in elections in other countries through disinformation campaigns aimed at undermining confidence in democratic processes, spreading conspiracy theories, and supporting divisive political candidates or causes. But the emergence of AI system has created more threats, as shown

[1] Podcast, Colin G érard, RFI, March 2024

by the social media influencing used by the company Cambridge Analytica[2] in the US 2016 elections [65]. It is indeed nowadays technically possible to differentiate between demographic and psychometric profiling techniques to influence political elections [36]. Demographic profiling is informational, segmenting voters based on factors like age, education, employment, and country of residence. Psychometric profiling is behavioral, allowing for voters' segmentation based on personality traits [65]. Another related trend is automatic content generation.

It is essential to critically evaluate information sources and be cautious of false or misleading narratives, especially in the context of online information consumption and social media engagement. Additionally, ongoing research and monitoring efforts by governments, think tanks, and civil society organizations appear important but not sufficient for identifying and countering disinformation campaigns effectively.

Several states and non-state actors or even groups or individuals can be threat actors regarding disinformation. A significant type of non-state actor is the so-called advanced persistent threat (APT), a term used to describe malicious, organized, and highly sophisticated cyber campaigns. APT groups are often funded by state governments, providing them with the resources to conduct cyber-attacks and other hybrid threats like disinformation [28, 47, 56]. These groups played a notable role during the Russian military invasion in Ukraine, acting as separate entities from the state despite government funding. Russian disinformation about NATO and the war in Ukraine achieves global reach through these non-state actors. Russian "influence-for-hire" firms, such as the Social Design Agency (SDA), the Institute for Internet Development, and Structure, have received substantial funding from Russia to spread disinformation. In response, the European Union imposed sanctions on SDA and Structure, recognizing these campaigns as threats to the EU and its member states [2].

Western democracies are increasingly challenged by narratives that exploit deep-seated fears and prejudices, fracturing societies along political, ethnic, gender, and religious lines [7, 9, 5, 49] and to amplify scapegoating approaches to create fear and anger. The scapegoating approach is indeed recognized as one of the strategies of manipulation to polarize the political debate [4, 6, 15, 31-33, 42]. From the rise of Euroscepticism, which culminated in Brexit, to the divisive rhetoric surrounding migration, gender, and religion, these narratives have often proven effective in manipulating public opinion and creating sharp societal divides. Identifying these polarizing narratives is crucial because they undermine social cohesion, fuel extremism, and threaten democratic stability. It can be used to undermine trust in institutions and in democratic systems, as it is here understood that Trust is an important pillar of stability for societies [15, 29, 30, 38-40, 46, 52, 55, 63, 64]. By recognizing and understanding these narratives, society can better counteract their harmful effects, prevent the spread of disinformation, and promote a more inclusive and unified European community.

When addressing the topic of disinformation, discussions often center on prevention strategies, its impact, and methods for detection. However, an equally critical aspect understands the threat actors responsible for disseminating disinformation. Since disinformation involves the deliberate spread of false information, the intent behind these actions is to craft and promote a deceptive narrative. By examining the actors who originated the disinformation, we can gain deeper insights into its mechanisms, and how to protect ourselves from it. This paper focuses on identifying the key threat actors involved in disinformation, focusing on who is posing significant risks to Europe.

The European Union (EU) classifies threat actors based on whether they are state actors, non-state actors, or proxies, and further categorizes threats by their attribution as either technical or political. When foreign entities engage in the dissemination of false information, it is referred to as "Foreign Information Manipulation and Interference" (FIMI). FIMI is characterized as a "mostly non-illegal pattern of behavior that threatens or has the potential to negatively impact values, procedures, and political processes." This activity is manipulative, intentional, and coordinated, involving both state and non-state actors, including their proxies operating inside and outside their territories [16, 6]. In a report assessing the cyber threat landscape, the EU Agency for Cybersecurity (ENISA) identified several primary motivations driving these actors. These motivations include geopolitical aims, intentions to cause disruption, ethical reasoning, and economic or financial gain [17, 12]. Having established that these actors operate with diverse motivations, we will now delve deeper into specific cases and types of threat actors.

As developed in this paragraph, disinformation can take different forms, processes, and threads, and can be relayed by different actors including states, groups or individuals, whereas the quick advancements in technology are amplifying the opportunities of targeted foreign information manipulation and interferences. The impacts can be manifold, and there is a legitimate amplifying concern among civil citizens in Europe. The following study of the online survey for AI4DEBUNK focuses on the perceptions on disinformation of civil society in different European countries.

3. Concerns over and Impacts of Disinformation: The Results of the AI4DEBUNK Online Survey

3.1. Methodology

The team of researchers for the EU funded project AI4DEBUNK project on AI and disinformation, created an

[2] Cambridge Analytica – an advertising company involved in the 2016 US campaign, amassed large amounts of data, built personality profiles for more than 100 million registered US voters and then, allegedly, used these profiles for targeted advertising.

online survey in June 2024, in the form of an online questionnaire, with a mix of multiple-choice questions and open ended questions. The questionnaire is available online[3]. It is anonymous and the names and email addresses have not been collected. The questionnaire was translated into French, Italian, Dutch, Latvian, Bulgarian, Ukrainian, German, Norwegian and Greek. It was published on social media and on polls online platforms. It was also disseminated to students for instance in Belgium and in the Netherlands. The poll is open until the end of 2024, and we are regularly collecting and analyzing the answers. We have collected 328 answers. The age groups are presented on figure 1.

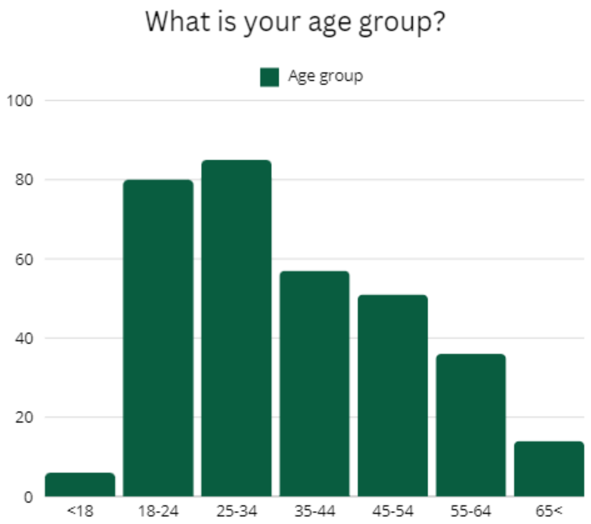

Figure 1. Age distribution of the online survey- Merged among the surveys in different languages.

Figure 1 illustrates the age distribution of all respondents, with a significant portion of answers coming from young adults (ages 18-24 and 25-34). The age groups: 35-44, and 45-54 showed more participation levels, while responses from older groups (65 and older) were notably fewer. As a limitation to the methodology, increased participation from the 55-64 and 65+ age groups could have made a more balanced representation across demographics.

The languages used in the survey do not perfectly represent specific countries. For example, English responses reflect a mix of various languages and networks. French is used in France and Belgium, and Dutch in Belgium and in the Netherlands. The German version was shared primarily through Poll-Pool, a platform popular with younger audiences, often used for academic research such as master's or PhD projects. Other languages were disseminated within specific networks, including WhatsApp groups and social media platforms. While some country-specific comparisons can be made, this is not intended to be a country comparison study. Rather, the multilingual approach was used to gather a broad range of responses from citizens across Europe.

Some of the findings are presented below to reflect further discussions in section 4.

3.2. Raising Concerns about Fake News

Regarding the question, "What impact do you believe fake news has on society?", there was a strong consensus among the respondents that it has a highly significant impact. In fact, 91% of all respondents indicated either a "significant impact" or "very significant impact" of fake news on society. This concern was shared across all countries, age groups, and genders. This highlights the broad recognition of fake news as a serious societal issue from a citizens' perspective.

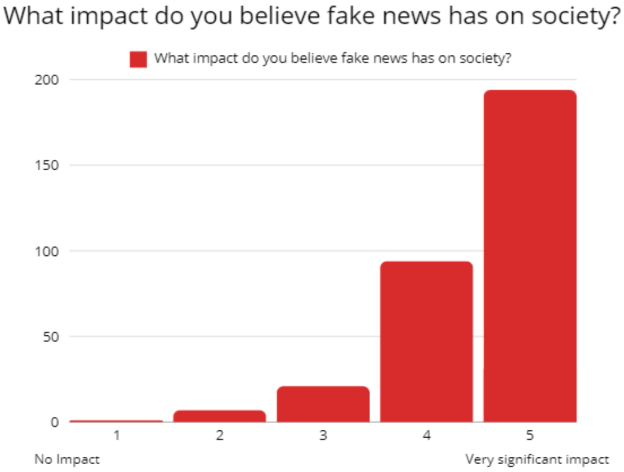

Figure 2. Answers from the respondents to the question 'what impact do you believe fake news has on society.?'

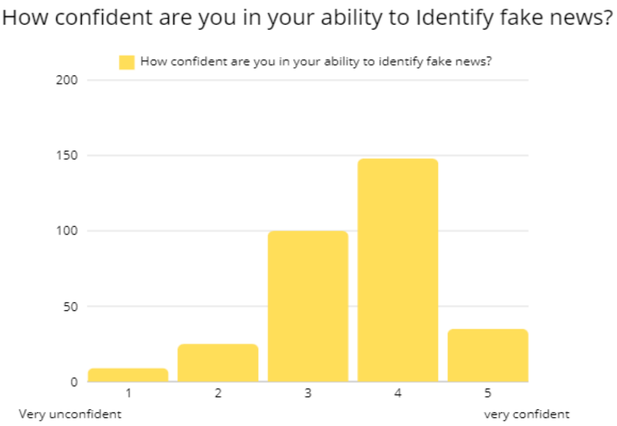

Figure 3. Answers to the survey respondents to the question 'how confident are you in your ability to identify fake news?" Merged from the surveys in different languages.

[3] https://docs.google.com/forms/d/e/1FAIpQLSfJ6RAs1makx1Y23CqKg2HZi5BuVtymJuiGvQ_ApO8jqJOwzQ/viewform?usp=sf_link

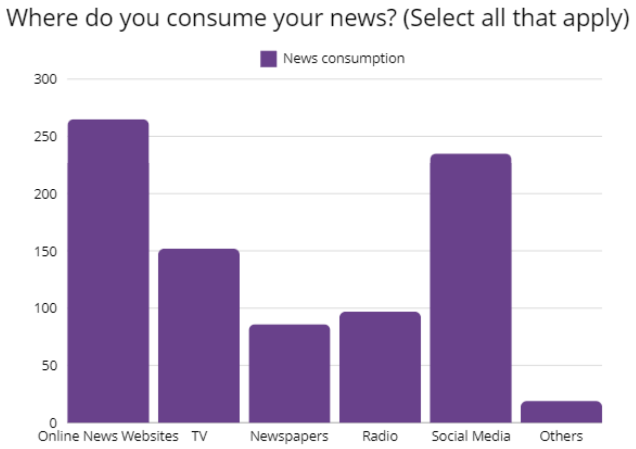

Figure 4. Main sources of information: answers from the respondents.

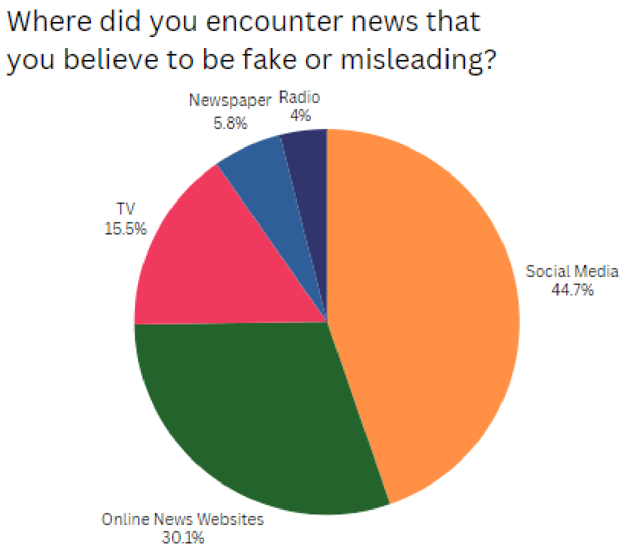

Figure 5. Answers to the question; where do you encounter news that you believe to be fake or misleading? Merged among the surveys in different languages.

There was a noticeable divide in respondents' confidence levels regarding their ability to detect false information. Most respondents reported feeling either somewhat “confident” or “neutral” in their abilities. Interestingly, younger people (aged 18-24) reported a higher level of confidence compared to older respondents. However, this varies a lot and is not a definite trend. Some of the languages related a bit higher confidence level, such as the Latvian survey, where 70.6% reported to either be “very confident” or “Confident” in their abilities to detect fake news, while this was only 40.9% for the French survey.

The figure above also shows that a majority of respondents reads news on online news website and on social media. Internet and social media represent therefore the main sources of ‘consuming’ news before television, radio and newspapers.

"Social media" was the most common source of fake news across all countries, with only minor variation. It was the top response in nearly every survey. "Online news websites" came in second, likely due to the rise of alternative or independent media outlets. Traditional media formats, such as TV, newspapers, and radio, which are more resource-intensive and dominated by established outlets, were less frequently cited as sources of misinformation.

However, there were some differences among the countries. In Latvia, respondents reported encountering fake news on "online news websites" just as frequently as on social media, the only country where social media was not the dominant source. This could be due to the prevalence of Russian-language media in Latvia, including online news websites linked to Kremlin-backed disinformation outlets. Similarly, in Germany, social media arrived first in the ranking, with 38 responses, while only 16 respondents pointed to online news websites. The age group could be an explanation.

In all surveys, TV was more frequently cited as a source of fake news than radio or newspapers, though TV appeared to be a particularly significant source in Greece. This could be linked to a polarized media landscape in Greece, where private media outlets can have political affiliations, contributing to the spread of misinformation on more traditional platforms like television. This phenomenon is however not limited to Greece.

In response to the questions about deepfakes, participants reported both their familiarity with the concept and their views on their potential societal impacts. The answers on familiarity were highly polarized: while many respondents indicated they were "very familiar" with deepfakes, a significant number reported little or no knowledge on the topic. Despite this divide, the majority of responders agreed on the potential consequences, with a clear consensus emerging that deepfakes could have a "significant" or "very significant impact" on society. Whether respondents were well-versed in the concept or not, most expressed concern about its possible effects.

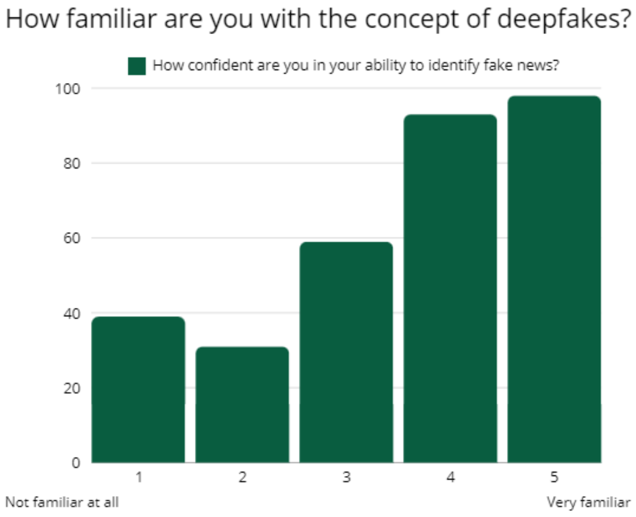

Figure 6. Displays the answers to the question: ‘how familiar are you with the concept of deep fakes?’. Merged from the surveys in different languages.

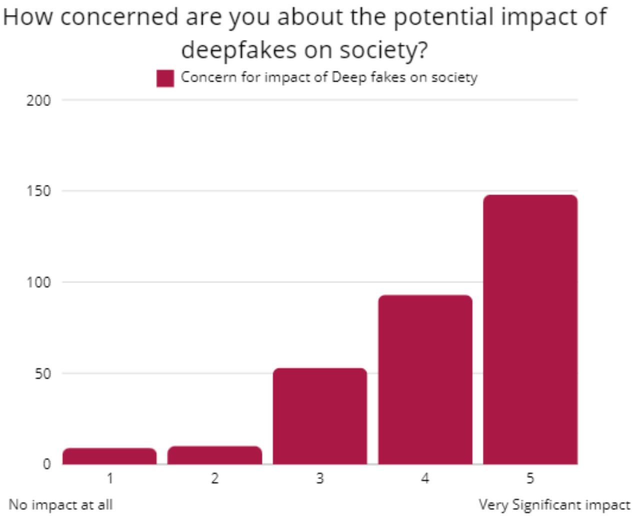

Figure 7. Shows the answers to the question 'how concerned are you about the potential impacts of deep fakes on society'.

Respondents were also asked to select two options they believed to be the most effective solutions for combating fake news. Two options stood out significantly more than the others: "Greater public education on media literacy" and "Increased collaboration between fact-checkers, journalists, and technology developers." Interestingly, these were also the preferred solutions of some field experts, suggesting that respondents demonstrate a solid understanding of the issue.

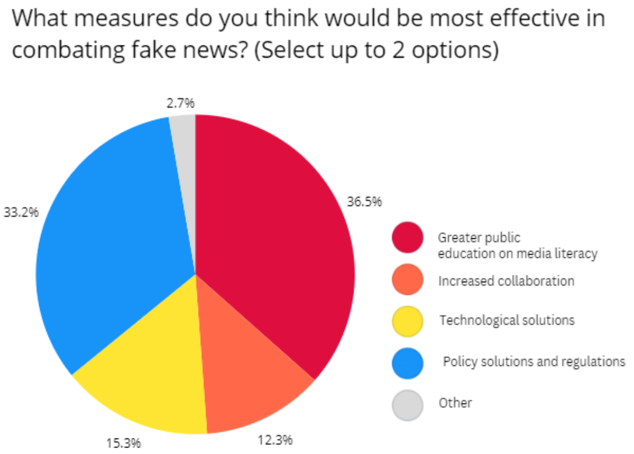

Figure 8. Displays the survey answers to the question: 'what measures do you think would be the most effective to combat fake news?'

Media literacy training as part of public education has proven highly effective in countries such as Finland and Estonia [3, 13, 44]. The educational system in these countries has successfully implemented strategies to counter misinformation by fostering resilience among their citizens, especially the younger generations. Other European countries are now attempting to develop similar solutions, and to equip the public with the skills to critically evaluate information. The online survey confirms that these attempts echo into a need among civil societies.

This trend of emphasizing media literacy and collaboration as key solutions was consistent across all languages and age groups, reinforcing the global recognition of these strategies as essential tools in the fight against fake news. This strong preference for media literacy and collaboration-based solutions reflects a growing awareness among the public about the multifaceted nature of fake news and the need for comprehensive responses. The alignment between experts' opinions and respondents' choices highlights a positive shift in public understanding, signaling that people are not only aware of the problem but also well-informed about potential solutions.

The survey brings interesting results against several assumptions: it confirms the prominent role of social media in being one of the main sources of information, and of disinformation. It highlights that there is a shared concern among the different European countries' civil societies about the potential impact of fake news. Even though young people are overrepresented in the survey, it also shows a discrepancy between the fears and concerns, and the understanding of the complex underlying mechanisms of making news viral through automated referencing or algorithms. It also shows a Trust gap in online information, which can decrease messages from scientific and public authorities. [58]

4. An Approach to Current EU Regulations

Stimulated by the lack of regulations in other Western democratic countries, the European Union (EU), followed by the national governments in the European Union have been actively developing policies and initiatives to tackle disinformation, particularly in the view of safeguarding democratic processes, protecting citizens, and promoting media literacy [55]. These policies and initiatives reflect the EU's commitment to addressing the multifaceted challenge of disinformation and protecting democratic values in the digital age. It diverts from other laws on disinformation like the one adopted in Malaysia or Singapore [60], as it does not interfere with preventive or reactive censorship except in the case of 'incitation to hate or crime'. These initiatives are strengthened by the approval of regulation on disinformation in most of the EU member states, which come in addition to the EU regulatory framework. This is important to note that the EU regulatory approach has also enabled a closer cooperation between the EU national governments in the area of cyber security, and joint cooperation to counter disinformation. Finally, we observe a struggle between the EU institutions and the main internet platforms to find the best way in tacking disinformation. Although the platforms submit a yearly plan, the question of regulation versus autoregulation is not over yet and it is expected that it will need to additional developments over the next years.

The paragraph below will highlight the main key regulations in the European Union, as the question has been one of the key important topics in EU policies since 2018. The European Commission launched the Code of Practice on Disinformation in 2018, jointly with different stakeholders including online platforms [18, 19]. Signatories committed to measures such as enhancing transparency in political advertising, disrupting fake accounts and bots, and empowering users to report misleading content. In December 2018, the European Commission published an Action Plan Against Disinformation, outlining measures to strengthen the EU's capabilities to counter disinformation campaigns [20-22, 24]. The plan includes initiatives to improve detection, analysis, and response to disinformation. It enhances coordination among EU institutions and national governments and promotes media literacy and critical thinking. The European Democracy Action Plan was put forward in December 2020 [23], to safeguard the integrity of elections and democratic processes in the EU. It includes measures to address disinformation, improve transparency of political advertising, support quality journalism, and strengthen media literacy.

The Digital Services Act (DSA) [26] was also proposed by the European Commission to update and harmonize rules for digital services in the EU. It includes provisions to tackle illegal content, including disinformation, by imposing obligations on online platforms to take measures to prevent the dissemination of harmful content while respecting fundamental rights. With the digital Service Act, the European Union transferred responsibility and accountability of moderation to the online platforms themselves.

From a prevention angle, the EU promotes media literacy initiatives to empower citizens with the skills to critically assess information and recognize disinformation. Funding programs support projects that enhance media literacy and promote quality journalism. The EU also established a Rapid Alert System in 2019 to facilitate the exchange of information among member states on disinformation campaigns targeting EU elections and other critical events. The system enables timely detection and response to disinformation threats. The European Digital Media Observatory (EDMO), launched in June 2020, was set up as a network of fact-checkers, researchers, and academics across Europe working to combat disinformation. It supports fact-checking activities, conducts research on disinformation trends, and provides analysis to policymakers and the public.

Finally, the AI Act [25], is the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a more visible role globally.

The AI Act aims to provide AI developers and deployers with clear requirements and obligations regarding specific uses of AI. The AI Act is part of a wider package of policy measures to support the development of trustworthy AI, while possible misuses have been identified by experts in recent years [61]. The AI Act ensures that Europeans can trust what AI has to offer. The risks of AI have indeed been identified as critical by most of the governments, but the EU is the first actor to regulate on the matter. Interestingly, the EU AI ACT proposes a model based on a pyramid, identifying the level of risks [67].

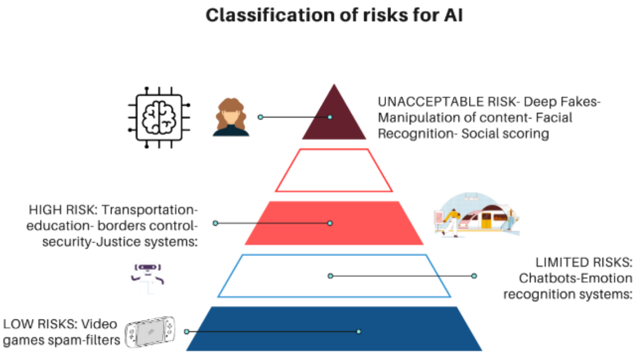
Classification of risks for AI
UNACCEPTABLE RISK- Deep Fakes- Manipulation of content- Facial Recognition- Social scoring
HIGH RISK: Transportation- education- borders control- security-Justice systems:
LIMITED RISKS: Chatbots-Emotion recognition systems:
LOW RISKS: Video games spam-filters

Figure 9*. Classification of risks for AI****.*** *Source Author-Inspired by [67].*

Limited risk refers to the risks associated with lack of transparency in AI usage. The AI Act introduces specific transparency obligations to ensure that humans are informed when necessary, fostering trust. For instance, when using AI systems such as chatbots, humans should be made aware that they are interacting with a machine so they can take an informed decision to continue or step back. Providers will also have to ensure that AI-generated content is identifiable. Indeed, AI-generated text published with the purpose of informing the public on matters of public interest must be labelled as artificially generated. This also applies to audio and video content constituting deep fakes.

Systems identified as high-risk include AI technology used in critical infrastructures (e.g. transport), that could put the life and health of citizens at risk; educational or vocational training, that may determine the access to education and professional course of someone’s life (e.g. scoring of exams), safety components of products (e.g. AI application in robot-assisted surgery); employment, management of workers and access to self-employment (e.g. CV-sorting software for recruitment procedures); essential private and public services (e.g. credit scoring denying citizens opportunity to obtain a loan); law enforcement that may interfere with people’s fundamental rights (e.g. evaluation of the reliability of evidence); migration, asylum and border control management (e.g. automated examination of visa applications); administration of justice and democratic processes (e.g. AI solutions to search for court rulings). As mentioned earlier, the display of generated AI images needs to be tagged with the mention ‘AI’ on any information published online, which should normally prevent the use and spread of ‘untagged’ deep fakes.

Most importantly, some uses with unacceptable risks will be banned: for instance, with the prohibition of real-time biometric identification by law enforcement authorities in publicly accessible spaces, but with some notable and clearly defined exceptions. Additional prohibitions include untargeted scraping of facial images for the purpose of creating or expanding facial recognition databases, emotion recognition but only at the workplace and in educational institutions (and with exceptions for safety and medical reasons), a very limited prohibition of biometric categorization based on certain specific beliefs or characteristics, as well as a limited and targeted ban on individual predictive policing. These bans are extremely important, as it should prevent targeted information manipulation, using AI.

The text has finally been approved in March 2024 by the EU institutions. It is important to note that the disinformation or malevolent intentions are not really part of the risk assessment, which rather tries to create a framework on future usages and automations. This AI Act is very much expected to provide a safe framework for the reasonable use of AI. It can, however, be expected that loopholes will be used, and that it will not be able to prevent malevolent uses. Some of the remaining challenges are detailed in section 5.

As shown by this paragraph, there have been active attempts to prevent disinformation since 2018. With the Digital Service Act, the responsibility of moderating the information is transferred entirely to the online platforms, while the AI Act is an attempt to regulate the use of AI not restricted to disinformation. Finally, the 2018 code of practice on disinformation could be a powerful tool, but it has not prevented an amplified spread of disinformation in the different European Countries as it mainly relies on self-regulation and reactive measures. This raises the question of the impacts of disinformation, the complexity of the threads and processes, and of the possible responses, which will be addressed in the discussion points below.

5. Discussion Points

There are remaining challenges in the current framework and in finding adequate responses to disinformation or misinformation, mainly explained by the complexity of processes, threads, mechanisms and shared responsibilities.

The current debates over social media and regulations overlook the role of malicious actors in spreading disinformation. Identifying these actors and tracing the origins of such threats is often incredibly difficult. Several studies developed by NATO Stratcom and the EU Disinfo Lab show that ‘inauthentic coordinated behavior’ is being used to spread fake news [53]. This makes it difficult to identify threads and to hold the authors accountable for their acts [54]. A general context of impunity may be created if no accountability mechanism is created.

Drawing from available literature and case studies, we tried to consider how Russian disinformation tactics have proven to be particularly adaptive. For instance, their strategy on platforms like TikTok focuses on attracting younger, liberal audiences with engaging content, gradually introducing propaganda after first building trust. This approach of embedding disinformation in narratives that resonate with specific demographics poses a growing threat, especially on platforms with weaker disinformation controls, like TikTok or Telegram. The fact that disinformation about the war in Ukraine was amplified by pro Kremlin media and channels was simultaneous to the spread of other ‘fake news’ related to climate change which authors are less easily identified. For this second case study (climate change), it is indeed more difficult to trace back ‘threads of disinformation’ and to identify malicious actors as less research is taking place. Further research in that area is needed.

There are also blurred lines between the ‘debunking’ of disinformation done by private actors, and the security agencies. Whereas citizens are encouraged to watch videos on how to recognize misinformation, and to make use of fact checkers manipulative disinformation campaigns are still rather unveiled by security agencies. France’s Viginum agency was set up in 2021 to detect digital interference from foreign entities aiming to influence public opinion. The agency reported to the media in February 2024 that they have uncovered more than 193 websites spreading disinformation directed through social media sites and messaging apps. According to the agency, the disinformation campaign was amplifying conspiracy theories and creating divisive narratives. It seems that even for security agencies, the characterization of the origin from the campaign is not always obvious, when the disinformation or misinformation needs to be traced back to foreign governments or to simple individuals acting in the interest of their country. In addition to this, hybrid warfare to which disinformation is only a tool, is combining cyberattacks with massive disinformation, creating risks for malevolent influence towards the media, the governments, the public infrastructure but also the civil society and the academic sectors.

Other responses like fact checkers also present limitations, as they are mainly responsible and not preventive. The number of fact-checkers around the world doubled over the past six years, with nearly 400 teams of journalists and researchers taking on political lies, hoaxes and other forms of misinformation in 105 countries[4]. Their expansion is however decreasing in numbers. It is difficult to assess their real impact for several reasons: Fact checkers may provide limited capacities: Most of the traditional media in Europe, but also the main platforms have created fact checkers. These fact checkers have yet however limited capacities to stop the spread of fake news and massive disinformation campaigns. They, however, play an important role in education towards rational thinking and highlight the need for civil society, the media and for citizens to ‘double check’ the information. Some campaigns like the ‘Matryoshka’ or ‘overload’ campaign in January 2024 have been targeting fact checkers to stop them, or to mock their impact.

Other research shows that ‘hostile actors persist in devising

[4] https://www.poynter.org/fact-checking/2022/391-global-fact-checking-outlets-slow-growth-2022/

innovative strategies to circumvent blocking and labelling mechanisms thereby effectively weaponizing the very measures designed to counteract information manipulation online [5, 9, 28, 47]. Independent journalism, able to cross-check information has never been as important as in recent years, while technology is jeopardizing their business models [37, 45, 50].

In addition to this, the difference and gradation levels between "disinformation" and "misinformation" is equally not entirely clear in the way to assess to what extent the facts distort reality, and to distinguish between opinions and fake news. False information exists on a spectrum ranging from unintentional misinformation to deliberate disinformation, as this is especially shown by the analysis of our case studies on climate change and the war in Ukraine. This point makes it arduous to draw a line between respecting freedom of speech, and countering deliberate disinformation.

The way information is framed, especially through the use of emotional language, can be seen as a factor that influences how readers interpret content. AI systems that can assess the emotional tone of articles already exist, yet caution was expressed about the risk of overgeneralizing different types of false information. Interestingly, the platform Google has created a 'prebunking' initiative aiming at empowering individuals to spot, prevent and detect disinformation online[5]. The platform helps to detect 11 manipulation techniques: Emotional language, false dichotomy, cherry picking, fake experts, red herring, scapegoating, ad hominem, polarization, impersonation, slippery slope, decontextualization. These tactics can indeed be used to identify disinformation, but not to prevent their large spread as automated cross referencing are the use of fake accounts or bots are used to spread out disinformation campaigns.

Given the complexity of determining the truth, fact-checkers continue to rely on human analyses, an approach also central. Suggestions as responses could include for media outlets to be notified when they publish incorrect information and a greater collaboration between journalists and fact-checkers to ensure the accuracy of news.

Another challenge will be to find quantitative indicators for fake news, and to go beyond an analysis which remains subjective in its interpretation. The elements for a 'fake news' to become viral, indeed depend on their interaction with a number of divisive matters appealing to curiosity or to specific emotions in a society, and this will be difficult to 'pre bunk' based on specific scientific algorithms. Personalized targeting, based on personal or psychological characteristics, can be combined with Natural Language Generation tools to create content for unique users. Moreover, the aggressive automated dissemination of disinformation during electoral campaigns can alter the perceptions of the political contexts, situations and challenges and orientate/manipulate the choice of voters.

Another impact, which should not be overlooked, is indeed the impact of targeted disinformation on election results [65]. For this, preventive, and not only reactive measures should be taken as responses.

'Relying on the collection and manipulation of users' data in order to anticipate and influence voters' political opinions and election results, user profiling and micro-targeting may pose a threat to democracy, public debate, and voters' choices' [36, 48]. This point is extremely important as 2024 has been an election year for half of the world's population, and the interferences are currently suspected but not visible as evidence is difficult to find, and the created tools, even including AI, are not yet able to debunk all interference.

Finally, while user profiling and political micro-targeting may seem like commercial advertising, these practices also pose issues regarding privacy and personal data protection. This takes us back to the underlying process of amassing and processing of vast amounts of personal data which is used in AI systems[6]. "Such data can be (…) stripped of its original purpose (s) and may be used for objectives the individual is largely unaware of – in this case, profiling and targeting with political messages – in contravention of existing EU data protection principles" [36].

In terms of response, and also although our survey (section 3) shows that there is a high level of confidence to identify disinformation, the increasing level of fake news is likely to have an impact by creating doubts and concerns among the readers, with the effect to dilute the information on true facts. There are difficulties and challenges in identifying up front what can be considered as trustworthy information and what is not. The results of our research show that while disinformation is spreading and is being used in increasingly complex and aggressive campaigns, the existing responses are only at an early stage. The current responses available are indeed 'reactive'. They rely on the social media platforms for moderation and on the users to become more 'critical' towards what they read. In addition to this, the authors of disinformation being located in foreign countries, the questions of 'accountability' and 'transparency' are limited, as it is complex for public authorities to locate the authors of disinformation campaigns and to hold them accountable. In this context the current regulations, although comprehensive, cannot be efficiently enforced.

6. Conclusions

The objective of this article is to bring some elements in order to build a first analysis on disinformation and misinformation, and the proposed responses highlighting how it is impacting society. It proposes a first set of analysis and theory framework based on existing literature, stakeholders' interviews, analysis of case studies and trusted reports. It goes through the processes, actors, impacts and narratives, and presents the results of our online survey. It also analyses the

[5] https://prebunking.withgoogle.com/

[6] And which will also be used in the AI related debunking systems.

current regulatory attempts by the European Union and especially the Digital Service Act and the AI act. Finally, it identifies the elements which still need to be further discussed before the approval of an evolutive theory framework. In particular the question of blurred lines between disinformation and misinformation, but also between cybersecurity and other types of responses proposed by media, social platforms and fact checkers. It also highlights a general context of impunity of malicious actors, even though impersonation or media spoofing exceed legal boundaries. It confirms that fake news and disinformation are also increasingly used as a political weapon in the democracies' political debates, highlighting the need for more research and for efficient preventive responses.

New technologies open new possibilities to create online communities and open opportunities to be collective, quick thinkers. But they also contribute to the spread of disinformation, misinformation and indirectly to a general breakdown of authority and an erosion of values (figures of authorities are replaced by endless access to information and connectivity including fake and manipulated content). Business leaders, community leaders and political leaders will need to cope with these changes. According to security experts, future generations of leaders will indeed face an utterly new environment that could be characterized by major trends: the generalization of uncertainties, the necessity of multi-tasking, the possibility to create new communities and the possibility to support more collective actions[7].

The digital revolution is certainly contributing to shaping a world of uncertainties: upheavals in the economy, politics, lifestyles. Young generations will need to accommodate these rapid changes, including by developing more IT skills, media literacy, competencies and understanding on how the algorithms are working. Leaders, self-employed workers, students and employees will need to develop new types of competences such as e-management, IT, communication, and to cope with unlimited instant information. Limits will however be deemed necessary to avoid technological changes such as the emergence of algorithms or bots influencing when not dictating lifestyles.

The online survey confirms that disinformation is a shared concern among civil society. The digital revolution—highlighted by AI but which began earlier—is one of the key factors shaping the world of tomorrow. The current findings show that responses to disinformation in European countries are not yet effective, despite the strong regulatory framework in place. As we attempted to demonstrate in this article, we are facing a massive shift in how we access information, while the existing responses to disinformation are still in their infancy. The article also shows an increased awareness in European countries about the impacts of disinformation. However, it reveals a gap between the ability to identify "fake news" and disinformation, and a limited understanding of the processes, threats, and actors involved in spreading disinformation.

Abbreviations

AI	Artificial Intelligence
EU	European Union
ENISA	European Union Agency for Cybersecurity

Acknowledgments

This project has received funding from the European Union's Horizon Innovation Actions under Grant Agreement No 101135757. Views and opinions expressed in this article are, however, those of the author(s) only and do not necessarily reflect those of the project partners, the European Union or the European Commission. Neither the European Union, the project partners nor the European Commission can be held responsible for these views and opinions. Contributors include Joen Martinsen, research assistant at Pilot4dev, who has coordinated the online survey, and the whole partners' team of AI4DEBUNK who contributed to the survey by circulating the questionnaire.

Author Contributions

Pascaline Gaborit is the sole author. The author read and approved the final manuscript.

Funding

Part of this research is funded by the project AI4DEBUNK funded by the program Horizon Europe under Grant Agreement No 101135757.

Data Availability Statement

More information is available on the following websites: www.pilot4dev.com, www.ai4debunk.eu.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] AI4DEBUNK 2024, 'Towards of Theory Framework', AI4debunk, Riga, 13 March 2024.

[2] Antoniuk, D. (2023, November 8). Russian 'influence-for-hire' firms spread propaganda in Latin America: US State Department. The Record by Recorded Future. https://therecord.media/russia-influence-for-hire-firms-latin-america-propaganda-us-state-department

[7] WIIS Women in international security, Brussels, 2016

[3] Art, S.: Media literacy and critical thinking. *International Journal of Media and Information Literacy*, *3*(2), 2018. 66-71. https://doi.org/10.13187/ijmil.2018.2.66

[4] Bauer M., Cahlíková J., Chytilová J., Roland G., Želinský T.: Shifting Punishment onto Minorities: Experimental Evidence of Scapegoating, The Economic Journal, Volume 133, Issue 652, May 2023, 1626–1640, https://doi.org/10.1093/ej/uead005

[5] Bergmanis-Kor āts G., Arhippainen M. et al. Virtual Manipulation Brief. Highjacking Reality. The increased role of Generative AI in Russian Propaganda. NATO Stratcom. 2024 https://stratcomcoe.org/publications/virtual-manipulation-brief-20241-hijacking-reality-the-increased-role-of-generative-ai-in-russian-propaganda/307

[6] Betz H. G., Oswald M. L. Emotional Mobilization: The affective Underpinnings of Right -Wing Populist Party Support., Palgrave Handbook of Populism 2021. 115-143. https://doi.org/10.1007/978-3-030-80803-7_7

[7] Bollmann, H. S., & Gibeon, G. (2022). The spread of hacked materials on Twitter: A threat to democracy? A case study of the 2017 Macron Leaks (Doctoral dissertation, Hertie School).

[8] Bontridder N. and Poullet Y. The role of artificial intelligence in disinformation. Data & Policy, 3: 2021, e32. https://doi.org/10.1017/dap.2021.20

[9] Butcher, P., & Neidhardt, A. H.: Fear and lying in the EU: Fighting disinformation on migration with alternative narratives. Foundation for European Progressive Studies, 2020. https://www.epc.eu/en/publications/Fear-and-lying-in-the-EU-Fighting-disinformation-on-migration-with-al~39a1e8

[10] Casten Stahl: On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective. Informing Science: The International Journal of an Emerging Transdiscipline Volume 9, 2006, 083-096 https://doi.org/10.28945/473

[11] Charillon F.. Guerres d'influence. Odile Jacob 2018.

[12] Cull. N. J., 2009, 'Public Diplomacy: lessons from the past' USC center of public diplomacy. https://digitallibrary.usc.edu/asset-management/2A3BF11FS2UH?FR_=1&W=1272&H=674

[13] Darwin Rusdin, D., Mukminatien, N., Suryati, N., Laksmi, E. D., & Marzuki: Critical thinking in the AI era: An exploration of EFL students' perceptions, benefits, and limitations. Cogent Education, 11(1), 2290342. 2024. https://doi.org/10.1080/2331186X.2023.2290342Page%205%20of%2018

[14] Dauksas V., Venclauskien éL., Urbanaviciut éK., Friedman O: War on all fronts: How the Kremlin's Media Ecosystem broadcasts the war in Ukraine. NATO Stratcom https://stratcomcoe.org/publications/war-on-all-fronts-how-the-kremlins-media-ecosystem-broadcasts-the-war-in-ukraine/301

[15] Deutsch M.: Trust and Suspicion, Conflict Resolution Number 2 (Vol8) 1958. https://doi.org/10.1177/002200275800200401

[16] ENISA European Union Agency for Cybersecurity (i), Lella, I., Ciobanu, C., Tsekmezoglou, E. (2023). ENISA threat landscape 2023: July 2022 to June 2023, (I. Lella, editor, C. Ciobanu, editor, M. Theocharidou, editor, E. Magonara, editor, A. Malatras, editor, R. Svetozarov Naydenov, editor, E. Tsekmezoglou, edito). Retrieved from: https://data.europa.eu/doi/10.2824/782573

[17] European Union Agency for Cybersecurity (ii), Tsekmezoglou, E., Lella, I., Malatras, A. et al., *ENISA threat landscape for DoS attack – January 2022 to August 2023*, European Union Agency for Cybersecurity, 2023, retrieved from: https://data.europa.eu/doi/10.2824/859909

[18] European Commission, 2018a, A Multi-dimensional Approach to Disinformation: Report of the Independent High-Level Group on Fake News and Online Disinformation. Directorate-General for Communication Networks, Content and Technology. Available at https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

[19] European Commission, 2018b, Code of Practice on Disinformation. Available at https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[20] European Commission, 2018c, Synopsis Report of the Public Consultation on Fake News and Online Disinformation. https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation

[21] European Commission (2018d) Tackling Online Disinformation: A European Approach (Communication) COM (2018) 236 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236

[22] European Commission (2020a) Assessment of the Code of Practice on Disinformation —Achievements and areas for further improvement. Commission Staff working document (SWD (2020) 180 final).

[23] European Commission (2020b) European Democracy Action Plan (Communication) COM (2020) 790 final. https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

[24] European Commission (2021) Guidance on Strengthening the Code of Practice on Disinformation (COM (2021) 262 final). https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation

[25] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[26] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) https://eur-lex.europa.eu/eli/reg/2022/2065/oj

[27] Fine G. A.: Rumor, Trust and Civil Society: Collective Memory and Cultures of Judgment. Diogenes 2007, 54 (1): 5-18. https://doi.org/ 10.1177/0392192107073432

[28] Foreign Affairs Review, 2017 'The meaning of sharp power: How authoritarian States project Influence', Foreign Affairs Review, 16th November 2017. https://www.ned.org/the-meaning-of-sharp-power-how-authoritarian-states-project-influence/

[29] Gaborit P.: Restaurer la confiance après un conflit civil, L'Harmattan 2009 a. https://www.editions-harmattan.fr/catalogue/livre/restaurer-la-confiance-apres-un-conflit-civil/45760

[30] Gaborit P.: La confiance après un conflit ou la confiance désenchantée, in Bertho A., Gaumont-Prat H. et Serry H. Colloque international La confiance et le conflit, Université Paris Vincennes Saint Denis 2009 b. https://www.libraires-ensemble.com/livre/1737783-colloque-international-la-confiance-et-le-conf--alain-bertho-helene-gaumont-prat-herve-serry-universite-paris-8-vincennes-saint-denis

[31] Girard R. The Scapegoat, Johns Hopkins University Press, 1986.

[32] Goodhart D.: The future to somewhere: The populist revolt and the future of politics. London, Hurst and Company, 2017. 9781849047999.

[33] G. Rodriguez-Pose. A.: The revenges of the places that don't matter- and what to do about it. Cambridge Journal of Regions, Economy and Society, II (I), 2017: 189-201. https://doi.org/10.1093/cjres/rsx024

[34] Grabner-Kräuter S.: Empiral Research in Online Trust. A Review and Critical Assessment. International Journal of Human-Computer Study. 2003 https://doi.org/10.1016/S1071-5819(03)00043-0

[35] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D.: Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 2019. 374-378. https://doi.org/10.1126/science.aau2706

[36] Kertysova K.: Artificial Intelligence and Disinformation How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered, security and human rights 29. 2018. 55-81. https://www.shrmonitor.org/assets/uploads/2019/11/SHRM-Kertysova.pdf

[37] Lloyd J. and Toogood L. (published with I. B. Tauris): Journalism and PR: News Media and Public Relations in the Digital Age. Oxford University and Reuteurs institute. 2015. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/Journalism%2520and%2520PR%2520-%2520News%2520Media%2520and%2520Public%2520Relations%2520in%2520the%2520Digital%2520Age_Extract.pdf

[38] Hamm, J. A., van der Werff, L., Osuna, A. I., Blomqvist, K., Blount-Hill, K. L., Gillespie, N., … Tomlinson, E. C.: Capturing the conversation of trust research. Journal of Trust Research, *14*(1), 1–7. 2024 https://doi.org/10.1080/21515581.2024.2331285

[39] Hardin R. (Ed): Trust and Trusworthiness. New York, Russel Sage foundation editions, collection on trust, volume 4, 2002.

[40] Hardin R. (Ed): Distrust, NYC, Russell Sage Foundation. 2004.

[41] Haiduchyk T., Shevtsov A., Bergmanis-Korāts G. AI in Precision Persuasion: Unveiling Tactics and Risks on Social Media. NATO Stratcom 2024 https://stratcomcoe.org/publications/ai-in-precision-persuasion-unveiling-tactics-and-risks-on-social-media/309

[42] Hersh M. A.: Barriers to ethical behaviour and stability: Stereotyping and scapegoating as pretexts for avoiding responsibility, Annual Reviews in Control, Volume 37, Issue 2, 2013, 365-381, https://doi.org/10.1016/j.arcontrol.2013.09.013

[43] King K., Wang b. Diffusion of real versus misinformation during a crisis event: A big data driven approach. International Journal of Information Management. 71. 2023 https://doi.org/10.1016/j.ijinfomgt.2021.102390

[44] Kueng L.: Hearts and Minds: Harnessing Leadership, Culture, and Talent to Really Go Digital, Oxford University, Reuteurs Institute, 2020.

[45] Kunelius R., Heikkilä H., Russell A. and Yagodin D. (eds) (published with I. B. Tauris):, Journalism and the NSA Revelations: Privacy, Security and the Press, 2017.

[46] Luhmann, N: Trust and Power: Two Works by Niklas Luhmann. Translation of German originals Vertrauen 1968 and Macht 1975. Chichester: John Wiley. 1979.

[47] Małecka, A. (2024). Non-State Actors in Nation-State Cyber Operations. *Rocznik Bezpieczeństwa Międzynarodowego*, *18*(1), 45-64. https://www.ceeol.com/search/article-detail?id=1255330

[48] Mont'Alverne C., Badrinathan S., Ross Arguedas A., Toff B., Fletcher R., and Kleis Nielsen R.: The Trust Gap: How and Why News on Digital Platforms Is Viewed More Sceptically Versus News in General, Reuters Institute, 2022 https://reutersinstitute.politics.ox.ac.uk/trust-gap-how-and-why-news-digital-platforms-viewed-more-sceptically-versus-news-general

[49] Moravcsik A.. Taking preferences seriously: A Liberal Theory of International politics', International Organization, vol 4, n°51, fall 1997, p 513-533. https://www.princeton.edu/~amoravcs/library/preferences.pdf

[50] Newman N.: Digital News Project: Journalism, Media and Technology: Trends and Prediction. Oxford University, Reuters Institute, 2024. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-01/Newman%20-%20Trends%20and%20Predictions%202024%20FINAL.pdf

[51] Persily N. and Tucker J. A: Social Media and Democracy The State of the Field, Prospects for Reform, Cambridge University Press, 2021. https://www.cambridge.org/core/books/social-media-and-democracy/E79E2BBF03C18C3A56A5CC393698F117

[52] Putnam R.: Making Democracy work: Civic traditions in Modern Italy, Princeton University Press 1993. https://doi.org/10.2307/j.ctt7s8r7

[53] Romero Vincente A et al.. ‘Coordinated Inauthentic Behavior’ EU Disinfo Lab 2024. https://www.disinfo.eu/publications/coordinated-inauthentic-behaviour-detection-tree/

[54] Sessa M. G., 2023, EU Disinfolab ‘Connecting the Disinformation Dots’ Friedrich Nauman Foundation. https://www.disinfo.eu/publications/connecting-the-disinformation-dots/

[55] Sessa M. G., Miguel R. The Doppelganger Case: Assessment of Platform Regulation on the EU Disinformation Environment. NATO Stratcom. 2024. https://stratcomcoe.org/publications/the-doppelganger-case-assess-ment-of-platform-regulation-on-the-eu-disinformation-environment/304

[56] Seligman A.: The problem of Trust, Princeton, Princeton University Press. 1997. 9780691050201.

[57] Shahbazi M., Bunker D. Social media Trust: Fighting misinformation in the time of crisis. Information Journal of Information Management. 77. 2024. https://doi.org/10.1016/j.ijinfomgt.2024.102780

[58] Six F. E.; Latusek D.: Distrust: A critical review exploring a universal distrust sequence, Journal of Trust Research, 13: 1, 1-23, 2024 https://doi.org/10.1080/21515581.2023.2184376

[59] Scheirer W. A Review of A History of Fake Things on the Internet. Stanford University Press 2023 http://www.sup.org/books/title/?id=35460

[60] Smith, R. B., Perry, M., & Smith, N. N.: Fake News’ in ASEAN: Legislative responses. Journal of ASEAN Studies, 9(2), 2021. 117-137. https:/doi.org/10.21512/jas.v9i2.7506

[61] Smuha, Nathalie A.. "Beyond the individual: governing AI’s societal harm". Internet Policy Review 10. 3 2021 https://doi.org/10.14763/2021.3.1574

[62] Stahl B. C., On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective’ Informing social science, Vol 9, 2006. https://doi.org/10.28945/473

[63] Sztompka P.: Trust a sociological theory, New York, Cambridge University Press. 2000. http://ndl.ethernet.edu.et/bitstream/123456789/17643/2/28.pdf

[64] Tilly C.: Trust and Rule, Cambridge University Press. 2005. https://doi.org/10.1017/CBO9780511618185

[65] Wade M.. Psychographics: The Behavioural Analysis That Helped Cambridge Analytica Know Voters’ Minds. The Conversation, March 21, 2018, https://theconversation.com/psychographics-the-behavioural-analy-sis-that-helped-cambridge-analytica-know-voters-minds-93675

[66] Whyte C. Deepfake news: AI-enabled disinformation as a multi-level public policy challenge, Journal of Cyber Policy, 5: 2, 2020; 199-217, https://doi.org/10.1080/23738871.2020.1797135

[67] Witzel L: 5 Things You Must Know Now About the Coming EU AI Regulation, https://medium.com/@loriaustex/5-things-you-must-know-now-about-the-coming-eu-ai-regulation-d2f8b4b2a4a9 2021 pp 128-146.