Handbook

on

Identifying and Countering Disinformation

Project: DOMINOES Digital cOMpetences INformatiOn EcoSystem
ID: 2021-1-RO01-KA220-HED-000031158

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the National Agency and Commission cannot be held responsible for any use which may be made of the information contained therein.

Introduction

The Handbook on Identifying and Countering Disinformation is the product of DOMINOES ERASMUS+ funded project, developed and implemented in a consortium with partners from Spain, Malta and Romania. The main goal of the project is to develop the citizens' abilities to recognise, evaluate and react appropriately to disinformation spread in the online environment.

The research contained in this handbook focuses on six aspects. The first chapter examines the current trends in the informational environment, the evolution of mainstream media and social media and how they influence the ways in which citizens gain access to information and the types of information they are exposed to. It also explores how narratives can be used in disinformation and propaganda campaigns, paying special attention to the rise in conspiracy theory dissemination. Another important issue that is investigated is the novel approach to employing intelligence in strategic communication as a means of setting the right frame of understanding for current societal evolutions.

The second chapter focuses on aggravating factors for the dissemination of disinformation such as individual and group factors, the role of influencers and pseudo-analysts, societal factors and technological factors so as to better understand how they interact in order to facilitate the spread of disinformation and to set the scene for analysing the best approaches to countering or limiting its effects and effectiveness.

The third chapter delves into the best known and most-widely employed methods to mitigate the dissemination of disinformation. Firstly, it examines the discursive, argumentative, narrative mechanisms that make disinformation attractive to audiences, and secondly, it explores the advantages and possible short-comings of critical thinking, media literacy, debunking, fact-checking and prebunking as the most extensively used and recommended means of countering disinformation.

The fourth chapter presents and evaluates the existing legal framework for countering disinformation and propaganda, focusing on how the legislation interacts with media freedom and the right to free speech, what role data protection can play in countering disinformation and providing case studies of how existing legislation is formulated in three countries: Spain, Malta and Romania.

The fifth chapter presents an overview of technological solutions, both existing and emerging, that could be employed to counter disinformation. The chapter introduces the technological solutions for spotting, flagging and removing disinformation as well as serious games solutions which are designed to prepare and train citizens to recognise it and thus limit its negative impact. The last section of the chapter explores the limitations that technology has with respect to identifying disinformation attempts.

The sixth chapter provides an analysis of public policies operating at the European level, as well as at a national level, in the three countries Spain, Malta and Romania, and explores their strong points as well as their limitations with respect to their potential effects on stopping the spread of disinformation. The chapter also provides a potentially innovative policy solution, by translating an approach so far employed in intelligence: the whole-of-society approach.

All the chapters in the *Handbook on Identifying and Countering Disinformation* are linked to a series of digital competencies extracted from Digital Competence Framework for Citizens (2022). The handbook represents a basis for the development of these competencies, by explaining what disinformation is, how it operates, what effects it produces, what measures can be taken against it, all the time providing clear examples, case studies, presenting lessons learnt and good practices, and also explicitly analysing what still needs to be done and what the main challenges in the fight against disinformation are. Thus, the handbook represents a first necessary step to developing societal resilience against this increasingly wide-spread phenomenon that attempts to shape, or better said, misshape, public discourse and debate, as well as democratic processes and societies.

RESEARCH TEAM

Dr. Cristina Ivan is a researcher in security and intelligence studies. She holds a PhD in cultural studies from the University of Bucharest, where she has researched violent religious extremism in cultural productions within the British discursive space (2000-2010). Over the past 15 years she has specialized in the cultural study of violence, radicalization and terrorism, propaganda and disinformation, critical intelligence studies, etc. She has taken an active part in European-funded projects targeting an enhanced understanding and early detection of radicalization, propaganda, disinformation, as well as designing both preventive and countering interventions.

Dr. Irena Chiru is a professor of intelligence studies at “Mihai Viteazul” National Intelligence Academy, Romania. Currently she serves as the chair of the International Association for Intelligence Education – European Chapter, a position from which she is advocating for building bridges between scholars and practitioners in intelligence. Her main research interests are intelligence analysis, critical intelligence studies, intelligence cultures, and strategic communication. She has extensive experience in managing European-funded projects on issues ranging from intelligence research, security issues, countering disinformation, enhancing civil society resilience to crisis situations.

Dr. Ruxandra Buluc is a senior researcher at the National Institute for Intelligence Studies in “Mihai Viteazul” National Intelligence Academy. Her main research interests are strategic communication, disinformation, foreign influence manipulation and interference, security culture. She works in European-funded projects which are aimed at building security culture and resilience to disinformation and radicalization.

Dr. Aitana Radu is the Security Research Coordinator within the Department of Information Policy & Governance. Her research focuses on different aspects of security science, from violent radicalisation to intelligence oversight. Since 2013, Dr Radu has carried out extensive EU-funded research focused on radicalization, law enforcement practices, the implementation of the European Investigation Order, developing security science (ESSENTIAL), fake news (DOMINOES) and intelligence analysis in the context of border security (MIRROR and CRITERIA projects). Dr. Radu obtained her M.A. (Comparative Political Science) from the University of Bucharest with a thesis on democratic transitions in the Middle East, her M.A. in the Management of Intelligence Activities for National Security from the National Intelligence Academy with a thesis on the security risks posed by the radical Islamic discourse, and her PhD in Intelligence and National Security from the National Intelligence Academy with a thesis on the transformation of intelligence organizations. More information and publications at https://www.um.edu.mt/maks/ipg/staff/aitanaradu.

Alexandra Anghel has been a young researcher at “Mihai Viteazul” National Intelligence Academy since 2015. She is pursuing a Ph.D. in Sociology, at the Doctoral School of Sociology, University of Bucharest with a thesis on the recruiting processes used by military institutions to attract young generations. She has published on intelligence and security studies and has been involved in several European-funded projects such as CARISMAND, ESSENTIAL, CITYCoP, THESEUS, ARMOuR etc.

Dr. Valentin Stoian-Iordache is a researcher with the National Intelligence Studies Institute. He is specialized in intelligence and security theory. He holds an M.A. and PhD in political science from the Central European University, Budapest and a B.A. in political science from the University of Bucharest. He has published work on topics such as hybrid warfare, the securitization of corruption, critical security studies and the ethics of intelligence.

Dr. Rubén Arcos is an associate professor (profesor contratado doctor) at the Faculty of Communication Sciences at University Rey Juan Carlos and member of the research group Cyber-imaginary. He serves as program Co-Chair of the Intelligence Studies Section at the International Studies association (ISA). His research is focused on intelligence services and intelligence analysis, foreign disinformation, and hybrid threats. He has been appointed national member for the NATO Science & Technology Organization’s exploratory group SAS-ET-FG “Prediction and Intelligence Analysis”.

Cristina M. Arribas is a researcher and Ph. candidate at the Faculty of Communication Sciences at University Rey Juan Carlos and member of the research group Cyber-imaginary. Her research is focused on disinformation, hybrid threats and Russian-China relations.

Ana Ćuća is a Research Support Officer at the Department of Information Policy & Governance at the University of Malta. Specialised in protection of refugees and criminalisation of humanitarian assistance, she is currently working on topics such as migration, security and disinformation. She holds a Master's degree in Human Rights with Clinical Specialisation from the Central European University and a Political Science degree from the University of Zagreb.

Kanchi Ganatra is a Research Support Officer at University of Malta's Department of Information Policy & Governance. Her day-to-day work focuses on EU-funded projects related to migration, sexual violence prevention, and disinformation. Her previous work involves ethnographic research among migrants in various parts of the EU - including Greece, Estonia, and Hungary. She holds an MA in Anthropology from Tallinn University and a BA in Social Psychology from the University of Mumbai.

Dr. Manuel Gertrudix is a professor of Digital Communication at the Rey Juan Carlos University, coordinator of the Cyberimaginary research group, and co-editor of the scientific journal Icono14. Specialist in digital communication and its application in several domains, he has been Principal investigator of 10 competitive national and international research projects. He is currently the Academic Director of the Master’s degree in journalistic research, new narratives, data, fact-checking and transparency at URJC-Fundación Maldita.

Dr. Ketan Modh is a Lecturer at the Department of Information Policy & Governance at the University of Malta. He received a double PhD as a Marie Curie Actions Fellow from the University of Groningen (The Netherlands) and the University of Malta. His current interests in teaching and research encompass identity systems, data protection and information law. He has experience working on multiple EU-funded projects in technology and law, covering topics that include surveillance and migration.

Cătălina Nastasiu is an associate teacher at the National University of Political Studies and Public Administration, Bucharest. Her areas of academic interest include the study of disinformation, the analysis of media frames and strategic narratives. She has professional experience in journalism, fact-checking and has been part of a series of initiatives to combat disinformation and promote media and digital literacy.

Contents

1. CURRENT TRENDS IN THE INFORMATIONAL ENVIRONMENT

Introduction

The first chapter of the DOMINOES Handbook sets out to map the main current developments in the informational environment so as to better understand and lay the foundation for the most efficient and effective means of countering disinformation. As such, the first section of the chapter identifies the evolution of disinformation in the mainstream media and social media and more precisely analyses the ways in which information which is not accurate finds itself in individuals' newsfeeds and social media pages, how and why it becomes viral and in what ways online and offline communication both contribute to the spread of disinformation. The second section focuses on the ways in which narratives could become the essence of disinformation and propaganda activities, by exploiting elements of cultural identity, ideology, feelings of anxiety and anger, etc. At the same time, narratives could provide the best instruments for countering disinformation and boosting resilience to propaganda. Related to narratives, but more ample in scope, conspiracy theories have also become an integral part of disinformation and propaganda campaigns, altering the ways in which social reality is construed by citizens and subverting constructive public debates. Last but not least, the chapter looks into a very current development in countering disinformation which is the instrumentalization of intelligence in strategic communication, whose goal is to prevent propaganda from reaching its goal. Strategic communication can convince the target audiences by informing them promptly and timely about the truthful evolution of events.

Digital competences addressed:

1.2 Evaluating data, information and digital content;
2.1 Interacting through digital technologies;
2.3 Engaging citizenship through digital technologies.

1.1 Evolution of disinformation in mainstream media and social media

Cătălina Nastasiu

Abstract

The present section investigates the ways in which disinformation is currently spreading in the online environment as well as in traditional media with a view to identifying the challenges that are raised by the increasing speed at which information (of any kind) circulates in society. It also provides an overview of the types of incorrect information that exist, their features and potential effects.

Main research questions addressed

- What is disinformation and fake news?
- What are the types of disinformation?

Evolution of disinformation in mainstream media and social media

Today, we are part of a community in a highly interconnected world, mainly by communication technologies. As we have become so interconnected, we share and look for our news online. The variety of information sources has contributed to the spread of alternative stories and facts, while the Internet and social media let us share any form of content. Fake news finds a fertile ground online and the information chaos of the pandemic has proven this statement. Most people don't fact-check or verify before sharing, therefore, they contaminate social media feeds with rumours, misinformation or conspiracy theories. In addition, social media filter algorithms are prioritizing the more familiar post to users; consequently, we are less exposed to alternative views. As social media has become a significant part of our life, it's crucial to develop some media literacy skills, including critical media literacy skills. During these times, digital skills education that will help us recognise fake news and take proper measures is a key competence for all, from an early age throughout life. In recent years, the issues surrounding the spread of disinformation in the online environment have been acknowledged and confirmed across the globe on several peak occasions, such as several electoral campaigns, Brexit, the independence referendum in Catalonia, the COVID-19 pandemic, the war in Ukraine. Its consequences, as documented in the literature, include negative effects on political attitudes, distrust in media, and polarisation of opinion within online echo-chambers.

The subject of "fake news" has become of global interest since 2016, with the result of the referendum on the exit of Great Britain from the European Union and with the result of the US presidential elections of the same year, especially since the label of fake news was one frequently used by former US president Donald Trump. Since then, the interest in information disorder or information pollution, as Clare Wardle and Hossein Derakhshan put it (2017, 5), in the major impact that digital platforms have on the information ecosystem, including on the traditional mass media that they have displaced from the position of information mediators, has grown constantly. Interest exploded during the pandemic, when the threat that misinformation could pose to public and individual health was strongly documented and acknowledged.

At the origin of the broader phenomenon of information disorder and the more precise one of disinformation are the changes produced by the explosion of digital platforms. In short, this online explosion has dislodged the mainstream media from the position of recognized intermediaries of information and allowed the production, dissemination and amplification of content without any editorial filter. Furthermore, the way digital platforms work has allowed amplification to happen not only organically (as a result of real people participating in online conversations), but also through algorithmic amplification and personalization (bombarding the user with content in accordance with their preferences deduced from previous digital behavior) and through artificial amplification (troll factories, bot factories, engagement amplification software, fake crowds, fake followers, etc.).

The phenomenon is not new, and it has been around for some time. When newspapers first appeared, hundreds of years ago, news and articles containing conspiracy theories, lies or sensational stories always sold well. *Fake News* became a catch-all description for the current information chaos. Fake news can be an invention, a lie, a media source that imitates an organization's official site or even a hoax created to persuade people that things that are unsupported by facts are true. The term "fake news" is often seen as inadequate and imprecise, while the concept of "disinformation" offers a broader perspective for the phenomenon. Disinformation may be defined as verifiably false or misleading information created, presented and disseminated for economic gain or to intentionally deceive the public. Disinformation includes various forms of misleading content such as fake news, hoaxes, lies, half-truths and, also, artificially inflated engagement based on automated accounts, *trolls*, *bots*, fake profiles that spread and amplify social media posts. Which is why understanding this phenomenon requires a complex approach.

What is fake news?

Fake news represents false or misleading information presented as real news. They can take the appearance of real news and mimic legitimate news sources very closely. These fake stories are deliberately fabricated to deceive and fool readers or to become viral on the Internet.

There are several warning signs that help us spot fake news. Firstly, we should consider and check the source, the author, and the date of the story. It is essential to verify the information, claims and cited sources by using a search engine, a reliable news source or a fact-checking organization. Many times, fake content comes from websites that use clickbait titles or stories, poor grammar and words in ALL CAPS or even websites that present past events as recent news and facts**.** Fake news also comes with a great emotional appeal and can provoke various emotional reactions, from fear and anger, to sadness and joy. In the age of technology and social media, fake news has more tools at its disposal than ever before, including text, video or photo manipulation.

Also, they can destroy trust, from trust in media and journalists to trust in public institutions, government, scientific experts or health experts.

This new phenomenon can be understood from a broader perspective than the buzz word "fake news" would suggest. This latter term is seen as "inadequate, imprecise and misleading", and the phenomenon requires a more inclusive and complex approach. For the purposes of this study, we adhere to the understanding of disinformation as "all forms of false, inaccurate, or misleading information" that was created to "intentionally cause public harm or for profit" (European Commission, 2018). When defining disinformation, the current focus is on the actual content and its truth value, and consequently, on specific countermeasures (i.e. fact-checking, debunking, coming up with counter-narratives), whereas digital disinformation relies on emotions and visual discourse, disseminated and, most importantly, amplified in the new digital ecosystem whose features we have previously underlined.

Online disinformation

The Internet and social media allow disinformation campaigns to be created immediately and through automated accounts, fake profiles, bots or "army of trolls" shared over digital platforms, while having the advantages of low cost, rapid spread and high impact. All of these actions and actors form an artificially inflated engagement (based on likes, comments, shares), that leads to the necessity to identify and combat the disinformation from a multi-layered perspective. Furthermore, the fact that manipulative and deceptive content manages to engage the users directly is a highly successful strategy, as it creates a sense of ownership over the message (users have the capacity to "endorse", contribute to, alter, and share disinformation that confirms their worldviews). This practice allows disinformation to infiltrate the most intimate spaces of communication. Given that viral content is inadvertently beneficial to digital platforms, their content curation algorithms are not prepared to deal with these particular challenges. Applying clear-cut rules and criteria for bringing down viral fake information would only tear up the whole fabric of social media, which is designed especially for promoting emotionally engaging content irrespective of its intention to deceive or not. Through artificial engagement, disinformation can reach large audiences, and has the potential to virally multiply its effects long before giving the digital platforms or public authorities the chance to spot it and react.

Fake or manipulative content has never been more prevalent than it is today. The Internet, new media and social media, along with instant messaging, are channels and means of communication through which the spread of misleading news and disinformation is facilitated. This phenomenon weakens democracy, lowers trust in the authorities and accentuates political and social polarization (also see 2.3). Furthermore, at the level of perception, disinformation is one of the major concerns of individuals, for example, 85% of European citizens consider that the existence of news that misrepresents reality or is even false is a problem for their country (Eurobarometer 2018), while 83% state that it is a problem for democracy in general.

The phenomenon of disinformation in the online environment takes on various forms. Researchers in the field include, within communication diseases *(information disorder*), various forms of manifestation of intentionally misleading information (*disinformation*) or unintentionally misleading information (*misinformation*). Some of these refer to content falsification, such as photo-video manipulation content, fabricated content (100% fake), impostor content (mimics legitimate sources of news and information). Others rather fall into the sphere of interpretations that may mislead: false or misleading contextualizations of facts, misleading content, clickbait (characterized by the lack of concordance between title and content, images and content, etc.), satire and parody (which through the effects documented on the audience may be harmful or misleading). The phenomenon of disinformation encompasses all forms of false, misleading or inaccurate content, from true information around which a false context has been generated, photo manipulation or video propaganda by actors with an ideological agenda, to entirely fabricated content.

Digital amplification

In the current digital context, disinformation coupled with technological possibilities has determined a technological revolution, producing what we could call "new generation disinformation", "disinformation 2.0". Amplification can be achieved through fake accounts, like factories, influencer networks (real or fabricated) and through "precision segmentation", "computational segmentation”, through which messages are targeted to users, taking into account their digital profile and digital fingerprint. Another important aspect of digital amplification is represented by the combination between bots and influencers. Lotito et al (2021) performed a study focusing on how three types of agents (a) commons (ordinary users); (b) influencers, (c) bots disseminate disinformation messages in open social networks (OSN) and the reach they each have. They reach two important conclusions: (1) bots, as automated accounts, help the conspirators promote their messages as well as keep them alive by constantly retransmitting them; (2) influencers “amplify the small world effect” meaning they speed up the spreading of information in general, which may also include fake news. Therefore, the authors argue that it is important to pay special attention to educating and correcting influencers as they are a vital node in the viralisation of disinformation, and, consequently, could play an equally big role in minimizing its spread and impact.

Disinformation can influence a number of social and political processes, from undermining democratic values and citizens' trust in governments, to polarizing discourse or amplifying divisions between different social groups or state political allies. Disinformation is created and circulated for many reasons, including political, financial, social or psychological reasons. For example, disinformation put into circulation for political reasons can generate political instability, influence citizens' behavior, undermine democracy or cause election fraud. For example, disinformation created for financial reasons, can have the goal to obtain income from web traffic and advertising, based on a business model centered on viralisation. Disinformation created or put into circulation for social and psychological reasons can be aimed at entertainment, the desire to harm known or unknown people on the Internet, the consolidation of group identity according to ideological orientation, political and ideological conviction/partisanship.

Understanding the communication flows between online and offline communication

Online communication has been identified as a great promoter of disinformation. However, one must not dismiss the role played by offline communication in the process. Communication flows are not limited to the online or to the offline, but often interact, enhance each other's reach, and, if used to their full potential, they could turn disinformation into a societal-altering instrument, whose effects become more and more visible in the real world.

Communication flows influence the reception of news stories and opinions disseminated in online media, and more broadly, they shape the opinions, attitudes, and behaviors of individuals particularly towards public issues. The COVID-19 pandemic has shown that perceptions of individuals about the measures implemented by public health authorities can have an impact on their acceptance or rejection of such measures and hence influence their behaviors and the global health situation. Misinformation and conspiracy theories disseminated in online and offline interactions have populated the pandemic context, generating confusion during the first stages and hence impacting public health.

There is a current trend to dismiss offline communication and face-to-face interactions of individuals in human social networks of family and friends. However, offline communication continues to be an important source of influence in shaping perceptions and behaviors in our current information environment of digital communication. Offline communication continues to be an important component of communication strategies, as demonstrated by studies on February 20 movement in Morocco,

> According to all the interviews with senior communication and political strategists of the movement, the activists had to rely on both online and offline communication platforms. They did not manifest as an oppositional binary. Instead, they functioned as an organic hybrid that combined face-to-face interactions and communicative practices on the ground, with online mobilization and exchange of information. In several ways the February 20th activists used complementary communication strategies for both the online and offline environments (Abadi 2015: 128).

Another important aspect which one should be mindful of when analysing information flows and online-offline interactions is the role that "opinion leaders" play in setting the tone for these debates and in framing and viralising the messages. In their classic study *The People's Choice*, Lazarsfeld, Berelson and Gaudet postulated a two-step communication flow and the importance of "opinion leaders" – in our digital era we may name them as "influencers"– in the reception of information by other less active media consumers,

> One of the functions of opinion leaders is to mediate between mass media and other people in their groups. It is assumed that individuals obtain their information directly from newspapers, radio, and other media. Our findings, however, did not bear this out. The majority of people acquired much of their information and many of their ideas through personal contacts with the opinion leaders in their groups. These latter individuals, in turn, exposed themselves relatively more than others to the mass media. The two-step flow of information is obvious practical importance to any study of propaganda (Lazarsfeld, Berelson and Gaudet 1948: xxiii).

Elihu Katz (1957), in a later review of the two-step flow hypothesis for *Public Opinion Quarterly*, highlighted three aspects of interpersonal relations in influencing decision-making: they are information channels, "sources of pressure to conform to the group's way of thinking", and also sources of social support (Katz 1957: 77).

The model has been subject to criticism since its formulation for different reasons (See: Weimann 2015), including the transformation of the information environment by developments in communication technology since the mass media era in which the hypothesis was formulated (Bennet and Manheim 2006). However, more recent studies have tested the hypothesis of the two-step flow in the context of Twitter-based discussion groups and found that a two-step flow of communication took place in the online political discussions (Choi 2014).

Also with respect to communication on online platforms, Hilbert et al (2017) developed a useful taxonomy of Twitter "communicator types" in the context of social movements in Chile – voice, media, amplifiers, and participants– and found that different proposed flow models –one-step, two-step, multistep– fit the finding of their empirical study (Hilbert et al. 2017). Along the same lines, Thorston and Wells have proposed the concept of "curated flows" of information/content, by considering that "the fundamental action of our media environment is curation: the production, selection, filtering, annotation, or framing of content" (2016, 310). In this framework they differentiate between five different actors practicing curation:

- Journalistic curation
- Strategic communication curation
- Individual media users
- Human social networks (family, friends, colleagues)
- Algorithmic filter.

These studies bring into sharper focus aspects related to how influencers may shape the informational environment and curate the information that is transmitted on social media (Ewing &Lambert, 2019). Recent studies and case analysis have focused on examining the ways in which social media influencers use digital media to shape public opinion, especially given their increasing involvement in societal debates, be they related to politics, education, healthcare, the environment, etc. Kadekova and Holienčinová (2018) identify four categories of social media influencers, according to areas of expertise: celebrities or macro-influencers, industry experts and thought leaders, bloggers and content creators, and micro-influencers.

Micro-influencers deserve increasing attention as they are more difficult to identify, however, their combined reach could prove transformative for society. Chen (2016) argue that given their smaller number of followers, micro-influencers are seen as being more authentic, closer to everyday people's concerns. Ong et al (2021) draw attention to the sway power that influencers can have, and they do not limit their research to what they term mega-influencers (people with more than 500,000 followers), but they also focus on micro (10,000-100,000 followers) and nano-influencers (1,000 to 10,000 followers). The reason behind this is that high profile mega influencers draw a lot of public attention to their posts, and this visibility exposes them to criticism, sanctions and even suspension of activity from social media when they are discovered as peddling disinformation. However, "micro-influencers roped in to seed political messages in a much more clever and undetectable manner. This trend, we argue, has grave implications for electoral integrity and creates challenges in enforcing democratic principles of transparency and accountability" (Ong et al, 2021). Micro-influencers are more likely to be able to infiltrate communities to which macro-influencers, perceived as too distant, or too high profile, or too controversial, would have no access. The reason behind this accessibility is what media anthropologists have coined "contrived authenticity" The term is used to "describe internet celebrities whose carefully calculated posts seek to give an impression of raw aesthetic, spontaneity and therefore relatability. This makes it easier for them to infiltrate organic communities and evade public monitoring" (Ong et al, 2021, 22). These micro and nano-influencers engage more directly with their followers and are more intimate in their approach to the dissemination of messages (be they fake or mot). For this reason, we argue, they are even more dangerous when it comes to disinformation, because they mimic the relationships that people have with their offline communities, close circle of friends, in which trust is a given, interests and points of view are shared freely and more often than not common, and therefore, fake messages could be more easily accepted, with less scrutiny and examination as they prey on the previously formed trust and commonly held beliefs.

The reason why offline interactions should never be neglected when analysing the reach and impact of disinformation is that those networks, in which people actually know each other well, have previously formed and established relationships, share a bond of trust, are fertile ground from the unconditional and unverified acceptance of fake content meant to mislead. And this dynamic, as previously explained, could be transposed in the online environment, in close communities, where people have a history of interactions and a set of common beliefs, which are similar to bonds of trust, and which lead to more permissiveness in terms of messages accepted.

Case study and lessons learnt - What are trolls and how do we avoid them?

Trolling is the behavior of posting content on the Internet, especially on social networks, with the express and sole purpose of provoking a reaction, sowing discord, provoking an emotional response. The most common behaviors that can be considered as trolling: personal attacks, provocative comments, insults or vulgar language. Often, the identity used for online trolling behavior is a fake one (fabricated social media account that cannot be associated with a real person).

What could alert us that an account is fake/fabricated?

1. The account is not associated with a real name, but with a pseudonym
2. The profile photo is not of a person, but of an object (flower, landscape) or an animal.
3. The presentation of the account does not provide credible information about the holder's personal/professional life (graduated school, job).
4. The posts do not refer to the owner's professional/personal life at all, they do not give any indication of the real existence of the owner.
5. Posts or comments associated with that account are trolling (posts on a single topic, provocative, inflammatory, vulgar language, personal attacks, grammatical errors, sloppy expression).

References:

1. Bennett, W. L., & Manheim, J. B. (2006). The One-Step Flow of Communication. The ANNALS of the American Academy of Political and Social Science, 608(1), 213–232. https://doi.org/10.1177/0002716206292266
2. Chen, Y. (2016, April 27). The rise of micro influencers on Instagram. Retrieved from: https://digiday.com/marketing/micro-influencers/
3. Choi, S. (2015). The Two-Step Flow of Communication in Twitter-Based Public Forums. Social Science Computer Review, 33(6), 696–711. https://doi.org/10.1177/0894439314556599
4. Erlich, Abadi, Houda, "The February 20th Movement Communication Strategies: Towards Participatory Politics." Dissertation, Georgia State University, 2015. doi: https://doi.org/10.57709/7367528
5. Funke, D. and Flamini, D., ‘A guide to misinformation actions around the world’, The Poynter Institute, 2018, available at https://www.poynter.org/ifcn/antimisinformation-actions/
6. Hilbert, M., Vásquez, J., Halpern, D., Valenzuela, S., & Arriagada, E. (2017). One Step, Two Step, Network Step? Complementary Perspectives on Communication Flows in Twittered Citizen Protests. Social Science Computer Review, 35(4), 444–461. https://doi.org/10.1177/0894439316639561
7. Kadekova, Z., & Holienčinová, M. (2018). Influencer marketing as a modern phenomenon creating a new frontier of virtual opportunities. Communication Today, 9(2), 90-105.
8. Katz, E. (1957). The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. The Public Opinion Quarterly, 21(1), 61–78. http://www.jstor.org/stable/2746790
9. Kjerstin Thorson, Chris Wells, Curated Flows: A Framework for Mapping Media Exposure in the Digital Age, Communication Theory, Volume 26, Issue 3, August 2016, Pages 309–328, https://doi.org/10.1111/comt.12087
10. Lazarsfeld, Paul F., Berelson, Bernard and Hazel Gaudet (1948). The People's Choice (2nd edition) New York: Columbia University Press.
11. Lotito, Q. F., Zanella, D., & Casari, P. (2021). Realistic aspects of simulation models for fake news epidemics over social networks. *Future Internet*, *13*(3), 76.
12. Michele Ewing, A. P. R., & Lambert, C. A. (2019). Listening in: Fostering influencer relationships to manage fake news. *Public Relations Journal*, *12*(4), 1-20.
13. Ong, J., Tapsell, R., & Curato, N. (2019). Tracking digital disinformation in the 2019 Philippine Midterm Election.
14. Thorson, K., & Wells, C. (2016). Curated flows: A framework for mapping media exposure in the digital age. *Communication Theory*, *26*(3), 309-328.
15. Törnberg, P., ‘Echo chambers and viral misinformation: Modeling fake news as complex contagion’, PLoS ONE, 13(9), 2018, pp. 1–21, doi: 10.1371/journal. pone.0203958.
16. Vosoughi, S., Mohsenvand, M.N. and Roy, D., ‘Rumor gauge: Predicting the veracity of rumors on Twitter’, ACM Transactions on Knowledge Discovery from Data, vol. 11, no. 4, 2017, pp. 1–36.
17. Weimann, Gabriel (2015). “Communication, Twostep flow,” International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Volume 4. Pp. 291-293. http://dx.doi.org/10.1016/B978-0-08-097086-8.95051-7
18. Understanding Information disorder https://firstdraftnews.org/long-form-article/understanding-information-disorder/
19. What are 'bots' and how can they spread fake news https://www.bbc.co.uk/bitesize/articles/zjhg47h
20. Tackling online disinformation https://digital-strategy.ec.europa.eu/en/policies/online-disinformation
21. INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c
22. Investigating Disinformation and Media Manipulation https://datajournalism.com/read/handbook/verification-3/investigating-disinformation-and-media-manipulation/investigating-disinformation

23. European Commission, ‘Final report of the High Level Expert Group on Fake News and Online Disinformation’, 12 March 2018, available at https://ec.europa. eu/digital-single-market/en/news/final-report-high-level-expert-group-fakenews-and-online-disinformation
24. Code of Practice on Disinformation (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation
25. Flash Eurobarometer 464: Fake News and Disinformation Online - Data Europa EU

1.2 Narratives and their use in disinformation and propaganda

Cristina Ivan

Abstract

Narratives refer to the stories that are told and the ways in which events and characters are framed in particular narrative contexts. Given their public appeal and their intrinsic focus on plot, they have become an instrument of choice for disinformation and propaganda. The present section explores the ways in which narratives can be weaponised by disinformation and propaganda campaigns and how and why the audience is captivated by them.

Main research questions addressed:

- What are narratives?
- How can narratives be employed in disinformation and propaganda?
- What functions do narratives play in disinformation and propaganda?

Narratives have been generally studied within the field of literary studies. For obvious reasons, they have been linked to art and literary works and associated with the cultural heritage of a society. A shift occurred in this perspective with the poststructuralists that, for the first time, highlighted the universal character of narratives which are articulated in a multitude of vehicles such as: spoken or written discourse, pictures, movies, gestures, graffiti, art and street performance etc. In this larger context, Roland Barthes insisted on narratives' "infinite variety of forms…present at all times, in all places and all societies" and on their universal character, as international, trans-historical and transcultural, started "with the very history of mankind" (Barthes and Duisit, 1975). Their universal character and fundamental function of articulating reality makes narratives in this broader social and cultural sense the main vehicle of identity formation and dissemination. Whether we refer to individual or collective identities, their sense making is inseparably linked to narratives. Hence, narratives' main function is that of representation and they are generally perceived as accurate reflections and expressions of what we see, how we are and what we cherish as individuals and communities. However, with postmodernism and poststructuralism, and especially in the works of Jacques Derrida and Michel Foucault, narratives were also revealed as inextricably linked to power formation and projection, hence acquiring social and political value. In the words of Somers, "it is through narrativity that we come to know, understand, and make sense of the social world, and it is through narratives and narrativity that we constitute our social identities" (Somers n.d., 606).

Furthermore, by the end of the 20th century and the beginning of the 21st, narratives have already made their way into social sciences, political philosophy, psychology, organisational theory or cultural anthropology and their understanding deepened into even more complex forms. Hence, from representations of reality, they came to be perceived as carriers of ontological meaning and linked to identity formation and vehicles of power projection. Narratives were associated to influence and manipulation, and last but not least, also perceived as powerful tools of community resilience. It is again Sommers who remarked that ''all of us come to be who we are (however ephemeral, multiple, and changing) by being located or locating ourselves (usually unconsciously) in social narratives rarely of our own making" (Somers n.d., 647).

The social implications of narratives reside in the fact that they are collectively created and embraced. Once paying allegiance to our values and traditions, one cannot evade the power of narratives that come to inhabit our mind-world and dictate the framework into which we understand reality. It is in this context that narratives have been employed by in information operations, being preferred tools of propaganda, disinformation and covert information manipulation.

Firstly, we shall decipher the use and functions of narratives in propaganda. When thinking of the stories used by the propaganda machine, one first must relate to the obvious persuasion function. All state propaganda is linked to conveying stories of legitimate power. Hence, from government sites to media state outlets, propaganda channels will tend to build stories of legitimacy, efficiency and purpose. However, the most obvious and openly declared propaganda is, the less efficient it turns out outside its close circles of followers. Adverse receptors will most likely tend to dismiss all propaganda that goes in favor of the adversary. Hence, the most effective propaganda is that which cannot and will not be attached to its agent and transmitter. And some of the most successful examples in history that illustrate the value of narratives in creating successful propaganda come from culture and have been revealed by the work of cultural historians. David Monod, for instance, discusses the case of the *Porgy and Bess* musical under the interpretation of the Gershwin Opera, that was used by the American State Department as a tool of propaganda in Europe in the 1950's. As Monod observes, the declared objective of the State Department announced 3 months in advance of the tour was "to counteract propaganda of two kinds related to the United States: First, that this country has no real culture, (or) native artists of creative vitality. Second, that the colored people have no opportunity to develop their abilities beyond a slave status". What this historical example hints at is that narratives circulated via fine artistic productions, beyond their esthetic value, can be instrumentalized in the propagandistic exchange of adversary states and in ways that appeal to the minds of the target country's citizens. "In short, Porgy was an opera which, while admittedly good art, had been re-fashioned by the Department of State to serve as even better propaganda" (Monod 2010, 2). The promotion of art as cultural diplomacy and outreach towards friend and adversarial countries alike is definitely not new and has been used many times in history before. Moreover, the use of culture as propaganda also hints to the conceptualization of the famous "soft power" term coined by Joseph Nye (1990).

Yet, as cultural historians often argue, cultural products will always bear a potential for subversion, their complexity making them ambivalent, unexpected interpretations assigned by different readers occurring at any time. A special part in creating powerful narratives has to be acknowledged especially in the case of performative arts, those that in the framing imposed by the director, can channel understanding of narratives in less ambivalent terms and in favor of a preferred interpretation - that being the case with music videos, theater plays, and especially movies. The "war on terror" master narrative created post 9/11 is another globally known example of how understanding of major historical events can be shaped not only by political discourse and strategic communication, but also in cultural productions reflecting the events - books, movies, TV series, documentaries, all in a diverse plethora of explanations that go beyond the event and into the making of history. And while the majority of cultural productions do not serve as propaganda tools, their significance can mainstream politically loaded statements and can attach a particular meaning to an entire epoch.

A comparative analysis of Monica Ali's novel Brick Lane, published in 2003 and the British drama film directed by Sarah Gavron launched in 2007, reflects the tension created between different viewers of the New York Twin Tower fall as a result of the 9/11 terrorist attack. Both underline the interplay of love and death, hope and revenge, whom are given precedent distinctively by the different characters and lead readers and viewers into searching for a more deep insight of jihadism and the underlying social and psychological grievances that trigger it. They will serve as an ambivalent narrative of terrorism as an extreme and violent response to social injustice and discrimination, while both works of art turn the spotlight on human agency and love as providing alternative personal pathways. Dozens of other film productions, from Zero Dark Thirty (2012) directed by Katryn Bigelow to 9/11: One Day in America (2021) individualize the story while taking a narrative that could easily shape perceptions of viewers worldwide.

While the most visible and easily recognizable, state media outlets and cultural productions are not the most powerful channels instrumentalized to create preferred significances to events. Narratives have in history been used also quite extensively in gray and black propaganda operations. And if high art can play a significant part in the soft power apparatus, one should not overlook the role played by popular art, by memes, podcasts, YouTube video channels, documentaries and mockumentaries, citizen journalism and any other form of collective, grass root formation of narratives as stories we tell each other about ourselves as individuals and communities.

Main challenges in addressing disinformation narratives

The main challenges in addressing propaganda and disinformation use of narratives is that authors and purposes of such narratives remain most often concealed and a specific link is difficult to trace. Illustrative examples available for study remain those offered by state affiliated channels. But to better understand the role of narratives, let us first define what propaganda and disinformation represent.

Propaganda

Propaganda represents the intensive manipulation of information to influence perceptions and the ability of the target audience to make objective decisions. The overall aim of propaganda is to obtain strategic advantages, political and financial capital brand and image promotion etc.

Propaganda was used historically in order to legitimate political regimes, advance certain ideological causes, and also as a way to mobilize the masses in case of armed conflict. For Lenin, propaganda had, at the beginning of the XX century, two main functions: to inform and mobilize own military troops and to undermine the morale and the trust of the opposite army forces (Lenin 2018). Lenin associated propaganda with the need to adapt the message to the historical, cultural, and social environment and with “agitation measures”. In order to understand the functions played by narratives in complex information operations though, propaganda and disinformation included, one needs to refer to historical examples, documented in the archives.

> Discussing the British – Indonesian – Malaysian confrontation in the 1960’s, which lead to the destruction of the Indonesian Communist Party, David Easter shows that enemy sides operated black radios, black newspapers, spreading of rumours and writing slogans on walls, simply to try and influence perceptions of adversarial sides (Easter, 2010, 9). The specificity of unattributable, disavowable or black propaganda is that it targets key segments of the target population (young army conscripts, students, minorities etc.) that could be influenced through subversion and psychological warfare to act in the benefit of the propaganda agent. And narratives again play an important part here, being the main leverage of influence and persuasion. More recent examples, such as the use of conspirational theories during the Covid 19 pandemic, also show this difficult attribution task which remains most of the time unaddressed.

While in white propaganda the producer of the material is clearly indicated, in the gray propaganda the producer remains unclear, and in the black propaganda it can be totally covert and the public deceived to belief in a fake author. (Nabb Research Center Online Exhibits, n.d.) (American foreign relations, n.d.) Nevertheless, all kinds of propaganda make use of narratives to persuade.

Disinformation

Building on the definition previously given to disinformation in section 1.1., we add that disinformation refers to an entire array of tactics and strategies used to propagate false, inexact or out of context information (therefore hijacked from their real meaning). Its intention is to provoke damages and/or profit. Continuous disinformation can severely affect democratic processes, national security, and social cohesion. In the long run, it undermines citizens’ trust in legitimate authorities, the democratic system and the benefits of the information society, thus diminishing citizens’ permeability to information, knowledge, and progress (see also 2.3). Hence, in this case too, the role of narratives is to politically loaded, subversive, aimed at creating multiple “truths” and hence sow distrust and confusion. Furthermore, studies dedicated to social media fake accounts and stories have shown that consumers tend to isolate in alternative realities, conspirational bubbles and post truth ecosystems, where an us vs them perception of reality is consolidated. Hence, disinformation in the social media age has become a powerful weapon waging war by manipulating perceptions (see also section 1.1).

In a public statement published on its official site, the Global Engagement center of the US State Department affirms that “Russia has operationalized the concept of perpetual adversarial competition in the information environment by encouraging the development of a disinformation and propaganda ecosystem. This ecosystem creates and spreads false narratives to strategically advance the Kremlin's policy goals. There is no subject off-limits to this firehose of falsehoods. Everything from human rights and environmental policy to assassinations and civilian-killing bombing campaigns are fair targets in Russia's malign playbook." (Global Engagement Center n.d.) Needless to say, Russian propaganda and disinformation are not the only ones using powerful narratives to advance strategic objectives. Other states, and especially autocratic regimes, have developed similar tools and means. However, the Russian ecosystem propaganda its narratives across the Internet and viralising content when needed to weaken the adversary is a telling example.

1.2.1 Case study (1) – The Russian Playbook

A comprehensive timeline of the first hundred days of the Russian war against Ukraine, drafted by EU vs. Disinfo shows a telling pattern of the fake narratives used in the information war. At the beginning of the aggression, mid-February 2022, the Russian propaganda machine advanced the idea that there is a Ukrainian crisis caused by the disregard of the West towards the "neo-Nazi crimes" of the Ukrainian government associated forces. Later on, EU and NATO were made responsible for the support of the Neo-Nazi movement, allegedly having organized a coup d'état to create a militarized Nazi state in Ukraine and install what was labeled as illegitimate government. Poland and Romania were accused of attempting to occupy part of Ukraine, while the Ukrainian military "was denounced" to be behind the actual Russian inflicted attacks (Kramatorsk station) or crimes (Bucha, Irpin etc.). In April 2022, US was already accused of operating biolabs to develop toxins that target the Slavic genotype while Ukraine was repeatedly accused of preparing attacks with biological, nuclear, or dirty bombs etc. (source https://euvsdisinfo.eu/uploads/2022/06/100-days-timeline-PDF.pdf).

The pattern shows a sequence of narrative that attempt to persuade audience by denying facts and blaming the opponent for own deeds: Russia did not attack and does not wage war, the Ukrainian population is decimated by its own government forces, attacks are inflicted by third parties and imagined enemies aspire to conquer Ukrainian territories. Denying the own crimes and justifying war through the existence of imaginary enemies seems to have been a preferred narrative themes used to confuse, detour attention or simply create as much as possible *plausible deniability*. By the end of 2022, Russian propaganda and disinformation seem to have focused on two main areas: (1) to control domestic audiences, maintain support and persuade that the Russian government is waging a just war against an imminent threat against Russian borders and identity and (2) to undermine support for Ukraine in European countries and the US.

Example

Kiev regime controlled by West, neo-Nazis, Lavrov says

- publication/media - tass.com
- Reported in: Issue 273
- DATE OF PUBLICATION: 25/02/2022

- Article language(s) English

> MOSCOW, February 25. /TASS/. The current Kiev authorities are being controlled by Western states led by the US, and by proponents of neo-Nazism, Russian Foreign Minister Sergey Lavrov said at a press conference Friday.
>
> "Nobody intends to attack the Ukrainian people; nobody intends to treat Ukrainian Armed Forces service in a manner that humiliates human dignity. We are talking about preventing the neo-Nazis and those who promote methods of genocide from ruling this country," he said. "Because the Kiev regime is currently subjected to two mechanisms of external control: the West, led by the US, and the neo-Nazis, who promote their 'culture,' which blooms in the modern Ukraine."
>
> Answering a question from an American reporter, Lavrov recommended the Western media to meticulously examine official statements made by Russia.
>
> "I drew the attention of two previous reporters to what President Putin said. I understand that you have other things to do than read the statements that describe the Russian position in minute details, but I invite you to do it nevertheless. Maybe, you will advise your Ukrainian colleagues, representatives of Ukrainian Armed Forces first and foremost, to read them."
>
> Russian President Vladimir Putin said in a televised address on Thursday morning that in response to a request by the heads of the Donbass republics he had made a decision to carry out a special military operation in order to protect people "who have been suffering from abuse and genocide by the Kiev regime for eight years." The Russian leader stressed that Moscow had no plans of occupying Ukrainian territories.[1]
>
> When clarifying the developments unfolding, the Russian Defense Ministry reassured that Russian troops are not targeting Ukrainian cities, but are limited to surgically striking and incapacitating Ukrainian military infrastructure. There are no threats whatsoever to the civilian population.

The key narratives used to systematically mock and devaluate Ukraine since 1991 have been also researched in a series of articles published by Inna Polianska[2], and listed below:

- Narrative #1: 'Ukraine is a failed state which never existed before the USSR's creation.'
- Narrative #2: 'Ukraine is not a sovereign state, but an "anti-Russia project" financed by the West to destabilise Russia.'
- Narrative #3: 'The Ukrainian language is an artificially created dialect of Russian with Polish influences.'
- Narrative #4: 'Ukraine is one of the most corrupt states in the world so it will never be ready for EU membership. Even Western weapons are stolen and sold to Russia.'
- Narrative #5: 'The Ukrainian government is not self-sufficient and is just following the instructions of Western leaders.'

[1] A History of Defamation: Key Russian Narratives on Ukrainian Sovereignty - EUvsDisinfo

[2] Idem

- Narrative #6: ‘Ukraine must be de-Nazified for infringing the rights of the Russian-speaking population and then integrated into Russia’.

A closer look at these narratives also shows that they can equally be flagged all across the ex-Soviet bloc, evidence from media content in e.g. Belarus, Moldova, Georgia, the Baltic countries or Romania showing similar opportunistic use of the 6 narratives every time the social and political context allowed it and if the historical background matched the potential use. This leads to the conclusion that what we face is rather a set of templates populated with updated content and used repetitively to create feelings of inferiority, distrust, confusion, fear.

In a similar vein, larger, more complex narratives have been employed to attack the liberal world and Europe in particular. Another EU vs. Disinfo article references the following adjacent storytelling frameworks:

- The elites vs. the people, a populist frame for numerous conspirational theories dedicated to Big Corporations, Jews, Muslims, Financial elite etc.
- Threatened values (and traditions o.n.) - a framework often used against minorities
- Lost sovereignty or threatened national identity
- Imminent collapse (of Europe)
- The hahaganda narrative (using sarcasm to annihilate evidenced accusations in e.g. the Skripal case)

1.2.2 Case study (2) - Analysis of the flows of information and disinformation which shape the conflict in Ukraine – Ruben Arcos

The Road to War: Propaganda and Disinformation Narratives

Since 2014 (Revolution of Dignity, annexation of Crimea and war in Donbas), there has been an increase in the pressure of propaganda and disinformation flows by the Russian Federation on Ukrainian territory. Russian state-controlled media and second-tier politicians focused on spreading narratives about the new government's collaboration with neo-Nazi and ultra-nationalist groups and the oppression towards Russian-speaking inhabitants.

After the annexation of Crimea and the success of the propaganda launched from Moscow-sponsored TV channels (Maschmeyer, 2021), the Ukrainian authorities became fully aware of the need to control their information space. In order to prevent Russians take control, the Ukrainian government ordered to close the Russian TV channels – In 2016 the number of closed channels amounted to 73- later in 2017 the measure was complemented by the blocking of Russian social networks, Vkontakte and Odnoklassniki (Gretskiy, 2022).

Due to these actions, it was possible to limit the penetration of the Kremlin's information influence campaigns within the Ukrainian space. Within this scope, there are logical differences in their impact according to the ethnolinguistic characteristics of Ukrainians, with greater acceptance of disinformation narratives among those with ties to Russia and barely noticeable among citizens without such connections (Ercher and Garner, 2022). On the other hand, there is evidence that the themes also affect the acceptance of malicious content, giving less credibility to those related to political, historical or military issues (Ibidem).

Narratives identified during the run-up to the invasion

The propaganda activity carried out since 2014 included statements from senior Kremlin officials mentioning narratives aimed at reinforcing Russia's aspirations over Ukrainian territory by creating a favourable breeding ground for attracting public opinion. These narratives became part of the agenda-setting of the official media apparatus of the Russian state but also began to spread from unofficial channels (foreign outlets, bots and trolls' networks, agents of influence…). According to Gretskiy (2022) the main narratives by chronology and context prior to the invasion of Ukraine

- 2014- Kyiv's new government is penetrated by ultranationalist and neo-Nazi groups. Ukrainian citizens with Russian ethnolinguistic ties are being oppressed and there is a plan to conduct a genocide of Russian inhabitants. Firstly, this narrative was disseminated by state-controlled media and second-tier politicians.
- 2018- After the new presidential re-election, Putin began to assert that eastern and southern Ukraine were "originally Russian".
- On 30 June 2021, for the first time, during his annual televised call-in show, Putin accused NATO of the "military incorporation" (военное освоение) of Ukraine (p.2) Previously, Lavrov and some military and special services top officials stated that the "expansion of NATO" to the CEE countries would lead to "military incorporation by the Alliance of the territory of new states" (p.2). After this Putin's statement, anti the NATO narrative was adopted by members of the government and high politicians. TV channels launched special programmes on NATO's "military incorporation of Ukraine", and state-owned online outlets ended each news report or expert commentary in Ukraine by mentioning Putin's statement on NATO.
- July 2021- In an article signed by Putin on the website of the Kremlin stated that Ukraine was not a true and sovereign state and again accused NATO of the military incorporation of the territory of Ukraine.
- October 2021- In an article published by Kommersant former Russian president Dmitry Medvedev lashed out at Volodomir Zelensky and accused him of weakness, absolute dependence and lack of principles. The article ended to warned that Russia would only negotiate with a pro-Russian president.
- 30 November during his intervention at the Russia Calling! Investment Forum, Putin referred to the Russian control in Donbas as "thus far unrecognised republics". This was an evident sign that Moscow would soon recognize its independency (p.3).

Propaganda, soft power and grey area

The Russian invasion of Ukraine supposes the consummation of the imperialist foreign policy that had been brewing since the late 2000s and that has been mediated by a process of historical revisionism in intimate relation to the doctrines of Near Abroad, the Ruskii Mir and the neo-Eurasianist theory of Alexander Dugin. Although there are different nuances between these formulations, all of them aim to expand the borders of the federation by annexing territories of the former USSR.

Justifying Russian foreign policy has been a central element of the Kremlin's propaganda and influence actions. Within its borders, since the second half of the 2000s, the Kremlin has been carried out the implementation of a patriotic policy (Kratochvíl and Shakhanova, 2020) aimed at favouring nationalism in the new Russia attracting the support of public opinion to the Putin regime and showing the West as the enemy of Russian people. The idea was to show the grandiosity of the achievements of the Putin regime but also of the Russian/Soviet tradition (Afanasiev, 2007 cited by Vázquez Liñan, 2010) where Putin is presented as a main guarantor following the steps of Catherine I or Petter I. It was also important to give continuity to the mythology of the Great Patriotic War forged in the Soviet period by which the USSR is presented as a guarantor of the free world, continuing with the great story of the defeat of Nazi Germany by the Red Army (Sherlock, 2016). Among the instruments used to materialize this policy, we must mention the revision of history textbooks, in which the West presented as antagonistic, the creation of several historical institutions -the Memorial, the Russian Academy of Sciences, and the Military Historical Society-, as well as the establishment of the Council for the Development of the Russian Film Industry financed by the state for the production of patriotic films where family values are promoted (Vázquez Liñan, 2010).

Abroad, numerous state-funded think tanks were created, notably the Russian Council on International Affairs (RSMD) and the Russian Institute for Strategic Studies (RISI, RISS) that joined the aim of the Gorchakov Foundation (Ministry of Foreign Affairs) in the performance of public diplomacy tasks. These soft power activities have taken place in parallel with other types of actions that are part of the framework of soviet active measures (Kux, 1985), such as the creation of fronts - some continuers of the Soviet period-, media outlets, agents of influence, or political parties and groups.

In relation to the information manipulation campaigns conducted by the Kremlin in the West, there is a large investigation focused mainly on the volume of messages, analysis of content, actors and technologies involved in production and dissemination. However, it is difficult to determine the real impact – behavioral changes – of a campaign in the absence of existing research in this field that works with real data. It is possible to venture greater confidence and credibility in these messages than that held by Ukrainians, also considering the ethnic-linguistic particularities of each country, its history, as well as geopolitical issues. This set of factors determines that populations are more or less receptive to malicious content since they are more or less sensitive to disinformation depending on the ties maintained with Moscow.

Information flows and disinformation: information treatment, characteristics and trends observed in the war in Ukraine

Since the start of the invasion on February 24, EUvsDisinfo has registered more than 5,000 pieces of disinformation targeting Ukraine. The narratives put forward to present a continuation with respect to the pre-conflict period – Ukraine is not a truly sovereign and independent state, NATO has carried out an "active military incorporation of the territory of Ukraine", the aim of the

West is to isolate Russia – along with others related to the course of events itself. The main novelty that occurs once the invasion began was that for the first time, the top official Russian explicitly called the Ukrainian government neo-Nazi (Maschmeyer, 2021). This direct attribution would be used to justify the invasion -see Putin's speech- and would connect with the narrative already used by the Soviets since the end of WWII to present the USSR as a bulwark of the struggle against fascism and defender of the free world (Luxmoore, 2019).

It is also important to mention the particularities existing in the selection of content and the informative treatment carried out by the Kremlin, its adaptation to the target audiences -external and internal-, as well as the selection of the dissemination channels. We can thus find different approaches to the treatment of the same event. Sometimes the same treatment of content can be given, although pursuing different objectives according to the public -e.g., presentation of images of Russian bombings to demoralize the Ukrainian side and on the other hand motivate the own audience-. Other times, the information is presented differently according to its audiences, for example, before the shipment of weapons to Ukraine by the West, the messages are contradictory, while the West is warned not to make shipments under the threat that the Kremlin will respond with nuclear weapons, the Russian population has transmitted the message that Western weapons are scrap, but it also takes advantage of the situation to encourage recruitment.

It has also been observed recently with the missile attack on Dnipro and other cities (January 15 and 16, 2023) how Russian TV has avoided broadcasting the images. As we have already pointed out, these images are used to encourage their own audience, although the version given is – as is the case – that it is a mistake of the Ukrainian forces that usually carry out attacks against their own targets, either for lack of professionalism or for carrying out false flag operations. Omitting the images is therefore incoherent and allows us to point more to express order, given the scope of the attack and number of victims, and the authorities don't want splashing across the media landscape Russia.[3]

On the alteration and construction of alternative versions of an event, it is worth mentioning the response of the Kremlin media apparatus on the Bucha Massacre (February 27-March 30, 2022), denying its responsibility for the events and accusing the Ukrainian side of having falsified the images. To confuse audiences Russians propagated on social media bad quality videos played in slow motion, giving the viewer the idea that the Ukrainians had created a fictional show and the corpses were actually actors playing a role. Similar explanations were given for the attack on Mariupol Maternity Hospital (9 March 2022).

Another trend observed during the conflict has been the decontextualization of content - a common technique in information influence operations - observed not only with images but also in content created from short videos of other events, something that has been seen in the social network TikTok. Likewise, Deep fakes technologies have contributed to the creation of false content, being notable a high-quality video where Zelenski appeared asking Ukrainian soldiers to surrender their weapons (Stănescu, 2022).

[3] https://euvsdisinfo.eu/shifting-the-focus-engineering-paranoia-manufacturing-fake-threats/ #

As a novelty not identified so far within the modalities of disinformation, videos have been observed explaining the steps that have been followed to debunk videos allegedly attributed to the Ukrainian side. This is another step-in disinformation since it is not about attributing one's own action to the opposite side, now the issue of disinformation is the discrediting itself. A phenomenon that Patrick Warren, a professor at Clemson University and co-director of the Media Forensics Hub, has referred to as a "false flag disinformation operation."

As for the social media ecosystem, there has been a boom in disinformation activity, configuring Telegram and TikTok as preferred networks. With the ban on RT and Sputnik within EU member countries, the presence of Russian politicians and military tiers increased, as well as embassies on social networks, mainly on Telegram. The choice of TikTok, on the other hand, is explained by its orientation to the short video.

Control of freedom of the press and expression

The Kremlin's control of freedom of the press and expression has contributed favourably to the spread of fake news and propaganda among domestic audiences but has also made it difficult for Western media and the public to obtain alternative versions to the official version manufactured by the Kremlin's propaganda apparatus.

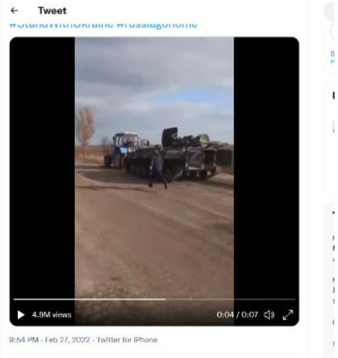
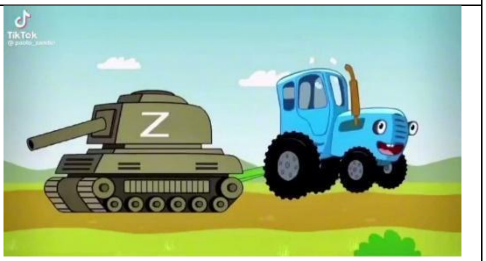
Russia enacted shortly after the invasion a new law restricting the freedom of speech and press that sets sentences of up to 15 years in prison for anyone who reports information not in accordance with the official position of the regime. According to the law, the use of the term war to refer to the conflict in Ukraine was prohibited, the term being the term of the choice of the special military operation. (Pavlik, 2022)

The new law led to the closure of the few remaining independent media outlets in Russia and the adaptation of Western media based there to the legal requirements imposed by the government. The Committee to Protect Journalists (2022) reports that at least 150 Russian journalists have fled in the aftermath of Putin's war on information. (Ibid.).

1.2.3 Case Study (3) - the Ukrainian Response to the Russian Playbook

In selecting our relevant case study material on inspiring practices using narratives to create resilience to propaganda and disinformation, we have opted for cultural productions that emerge out of the participatory digital culture, where creation is often anonymized, while co-production and non-attribution are widely shared behaviors. At the same time transgressive and empowering, these cultural productions derive their force from the amount of user interaction generated and the real time meaning making process they foster and encourage with digital users. Most rely on a media account of real life events only to then transgress into the symbolic regime and start generate meaning(s) by the engagement of the audience. The reason behind this choice is that within the larger framework of cultural productions, these are the ones that record changes while in the making and offer a great opportunity to the researcher to observe grass root shifts in the social world. Such environments offer most prolific X-rays of every day communication and interaction within communities, also linking up individuals at a global scale.

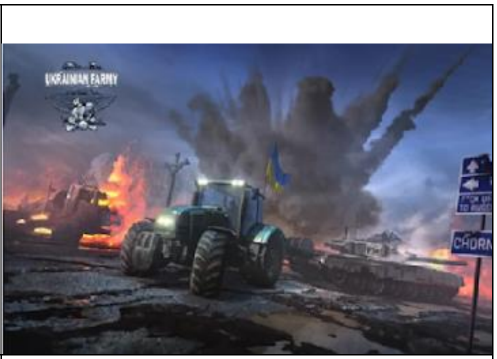
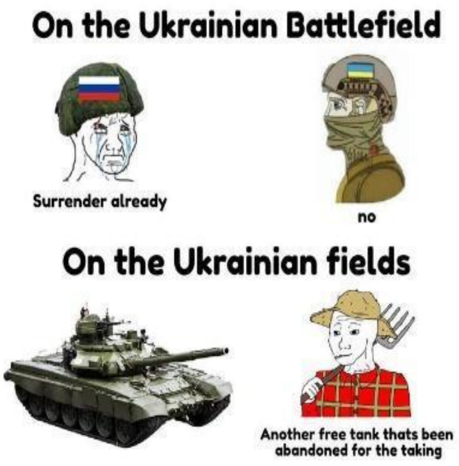
In the image – text analysis we shall focus our attention on the language, the interpretative frames that build up on the primary event and its manifold constructed significances, and the cultural repertoire used to translate experience into a viable, forceful meaning. Finally, by making appeal to archetypal analysis, we shall attempt to prove that collectively produced stories engage the full force of a protective factor and create mechanisms of resilience at community level. We shall analyse images, memes, videos games and cartoons in order to uncover the ways in which grassroot resilience to disinformation is constructed through the narratives they put forth.

As early as March 2022, the first video[4] from Ukrainians had appeared online (Facebook, Twitter, TikTok). It portrays a blue tractor which is towing a Russian tank, a reflection of an actual event that took place. It was accompanies by hashtags that highlighted Ukrainian resistance and toughness #StandWithUkraine and #russiagohome.	 Tweet 4.9M views 0:04 / 0:07 9:54 PM · Feb 27, 2022 · Twitter for iPhone
The narrative created and the associations with Ukrainian resistance endured, and at the end of July 2022, a google search of "blue tractor" + "Ukraine" yielded both text results as well as cartoons[5] and promotional materials depicting the already famous image. The cartoon featured here resembles a character in the famous kids' movie Cars. The artist[6] Синий Трактор (Blue Tractor) apparently took over the tractor from a popular children's show and adapted it to the current war reality, while, at the same time, showing Ukrainian courage and resilience against Russian aggression.	

[4] https://twitter.com/olex_scherba/status/1498023662695419910

[5] Viral cartoon shows Ukrainian tractor dragging Russian tank: What's the story behind it? | Euronews

[6] https://www.youtube.com/watch?v=henlXJKjasc

<table>
<tr>
<td>A video game portrays the same event, a Russian tank captured and towed by a Ukrainian blue tractor, this time armoured, depicted as one of the scariest combatants on the battlefield. Thus, in a David meet Goliath type of script, bravery becomes the new normal, and even a humble tractor (which can be associated to any, ordinary person) can become a hero of anti-Russian resistance.</td>
<td>
</td>
</tr>
<tr>
<td>Memes have become an integral part of the information operations in Ukraine since the Russian invasion. Plenty of memes were created in reference to the blue tractor narrative. As explained, “memes about Ukrainian farmers stealing tanks surfaced on multiple social media platforms like TikTok and Instagram, going into early March 2022, referencing other memes like Devious Licks, Ben Affleck Smoking and But It's Honest Work, among others. However, most Ukrainian farmer memes were shared on Reddit.”[7] The contrast that the meme we selected exhibits between the tearful call for a surrender and the fact that Russian tank is abandoned only goes to show that the Ukrainians are resilient, determined and calmly wait for the Russians to make mistakes (such as abandoning a functioning tank).</td>
<td>
On the Ukrainian Battlefield
Surrender already
no
On the Ukrainian fields
Another free tank thats been abandoned for the taking

</td>
</tr>
<tr>
<td>Postage stamps have also been issued in July 2022[8]reflecting the blue tractor narrative. The stamp design evokes the famous event, and depicts it in the national Ukrainian colours.</td>
<td> </td>
</tr>
</table>

Table 1. Analysis of memes used by Ukraine is response to Russian propaganda

[7] https://knowyourmeme.com/memes/ukrainian-farmers-vs-russian-army

[8] https://www.linns.com/news/world-stamps-postal-history/winning-design-in-ukraine-s-second-design-contest-features-tractor-and-tank

In conclusion, narratives can be used as a potent tool not only for the proliferation and weaponisation of propaganda and disinformation, but also as resilience-building instruments. Propaganda narratives reflect the concerns, weaknesses, insecurities of a target audience and as such could be used to enhance feelings of anxiety, anger, hopelessness, resentment. However, narratives can also empower, give hope, unite communities around their message, based on common values, interests, desires and dreams. It is a matter of understanding the cultural characteristics of various communities in order to create meaningful and attractive narratives that would prevent disinformation and propaganda from taking hold in those environments. The case study analysis of the resilience-building narratives is a brief insight into the power that narratives have which is both inspiring as well as restorative and cathartic.

References:

1. American foreign relations. n.d. "Propaganda - Types of propaganda." https://www.americanforeignrelations.com/O-W/Propaganda-Types-of-propaganda.html.
2. Aro, Jessika. (2016). "The cyber-space war: propaganda and trolling as warfare tools." European view 121-132. doi:10.1007/s12290-016-0395-5.
3. Barthes, Roland, and Lionel Duisit. (1975). "An Introduction to the Structural Analysis of Narrative" An Introduction to the Structural Analysis of Narrative (The Johns Hopkins University Press) 237-272. http://www.jstor.org/stable/468419.
4. Baumann, Mario. (2020). "Propaganda Fights' and 'Disinformation Campaigns': the discourse on information warfare in Russia-West relations." Contemporary Politics (Routledge) 1-20. doi:https://doi.org/10.1080/13569775.2020.1728612.
5. Best, Shivali. (2017). "The spread of fake news on Facebook and Twitter is made worse by social network algorithms." Mail Online, iunie 20. http://www.dailymail.co.uk/sciencetech/article-4621094/Are-Facebook-Twitter-ENCOURAGING-fake-news.html.
6. Bonino, Silvia, Elena Cattelino, and Silvia Ciairano. (2003). Adolescents and Risk Behaviour, Functions and Portective Factors. Torino: Springer.
7. Bradshaw, Samatha, and Philip P. Howard. (2018). "Why does Junk News Spread so Quickly across Social Media? Algorythms, Advertising, and Exposure in Public Life." http://comprop.oii.ox.ac.uk/research/working-papers/why-does-junk-news-spread-so-quickly-across-social-media/.
8. Bradsma, Bart. n.d. Inside Polarisation . https://insidepolarisation.nl/en/.
9. Candaele, Kelly. (2020). "Coronavirus is a political problem, not just a health problem. Remember that when you vote." The Guradian, March. https://www.theguardian.com/commentisfree/2020/mar/19/coronavirus-political-problem-health-voting-elections.
10. Chekinov, S.G., and S.A. Bogdanov. n.d. "The Nature and Content of a New-Generation War." Military Thought. http://www.eastviewpress.com/Files/MT_FROM%20THE%20CURRENT%20ISSUE_No.4_2013.pdf.
11. Culloty, Eileen, and Jane Suiter. (2021). Disinformation and Manipulation in Digital Media. Routledge .
12. "Disinformation: how to recognise and tackle Covid-19 myths ." News, European Parliament . 30 March 2020. https://www.europarl.europa.eu/news/en/headlines/society/20200326STO75917/disinformation-how-to-recognise-and-tackle-covid-19-myths.
13. Easter, David. (2010). "British Intelligence and Propaganda during the 'Confrontation', 1963-1966." Intelligence and National Security 1-21.

14. Edelman's Trust Barometer, Trust Inequality. n.d. "Edelman." http://edelman.edelman1.netdna-cdn.com/assets/uploads/2016/01/2016-Edelman-Trust-Barometer-Global-_-Mounting-Trust-Inequality.pdf.
15. Edward, Herman, and Noam Chomsky. n.d. "A Propaganda Model." In Manufacturing Consent, by Herman Edward and Noam Chomsky.
16. Edward, Lucas, and Pter Pomeranzev. (2016). Winning the Information War. Center for European Policy Analysis .
17. "EEAS SPECIAL REPORT UPDATE: Short Assessment of Narratives and Disinformation Around the COVID-19 Pandemic." EU vs Dinsinfo. April 01. https://euvsdisinfo.eu/eeas-special-report-update-short-assessment-of-narratives-and-disinformation-around-the-covid-19-pandemic/.
18. Erlich, Aaron and Garner, Aaron (2023). Is pro-Kremlin Disinformation Effective? Evidence from Ukraine, The International Journal of Press/Politics, 28(1) 5–28 https://doi.org/10.1177/19401612211045221
19. EU vs. Disinfo. (2016). "Estonia is building a concentration camp for its Russian-speaking citizens." EU vs Disinfo. https://euvsdisinfo.eu/report/estonia-is-building-a-concentration-camp-for-its-russian-speaking-citizens/.
20. EUvsDisinfo (2023) Shifting the focus, engineering paranoia, manufacturing fake threats https://euvsdisinfo.eu/shifting-the-focus-engineering-paranoia-manufacturing-fake-threats/# Cambiar el enfoque, paranoia de ingeniería, fabricar amenazas falsas - EUvsDisinfo [Consulted: 19/01/2023]
21. Farmy, Ukrainian. n.d. https://ukrainian.itch.io/ukrainian-farmy.
22. Foresman, Galen A., Peter S. Fosl, and Jamie Carlin Watson. (2017). The Critical Thinking Toolkit. Wiley Blackwell.
23. Gertrudis-Casado, María-del-Carmen, María-del-Carmen Gálvez-de-la-Cuesta, Juan Romero-Luis, and Manuel Gértrudix Barrio. (2022). "Los serious games como estrategia eficiente para la comunicación científica en la pandemia de la Covid-19." Revista Latina de Comunicacion Social. doi:https://doi.org/10.4185/RLCS-2022-1788.
24. Global Engagement Center. n.d. Disarming Disinformation: Our Shared Responsibility . https://www.state.gov/disarming-disinformation/.
25. Gray, Ann. (2003). Research practice for cultural studies. Ethnographic methods and lived cultures. London: Sage Publications.
26. Grejdeanu, Tamra. (2017). "Propaganda rusă în Moldova. Cum funcționează?" Radio Europa Liberă. aprilie 28. Accessed iulie 30, 2018. https://www.europalibera.org/a/propaganda-rusa-in-moldova/28457231.html.
27. Gretskiy, Igor (2022) Russia's Propaganda War, Russia's War in Ukraine Series No. 9, Aug, 2022
28. Guess, A. M., and B. A. Lyons. (2022) "Misinformation, Disinformation, and Online Propaganda. Social Media and Democracy, 10–33. doi:10.1017/9781108890960.003."
29. Helmus, Baron, Radin, Magnuson, Mendelsohn, Marcellino, Bega, Winkelman. (2018). Russian Social Media Influence. Understanding Russian Porpaganda in Eastern Europe. Rand Corporation.
30. Hicks-Goldston, C. (2019). The new digital divide: Disinformation and media literacy in the US. Media Literacy and Academic Research, 2(1), 49-60.
31. Hinchchliffe, Tim. (2020). "Exposing echo chambers to eradicate the plague of propaganda." The Sociable. https://sociable.co/social-media/exposing-echo-chambers-to-eradicate-the-plague-of-propaganda/.
32. Humprecht, Edda, Frank Esser, and Peter Van Aelst. (2020). "Resilience to Online Disinformation: A Framework for Cross-National Comparative Research." The International Journal of Press/Politics 1-24.
33. Humprecht, E., Esser, F., Aelst, P. V., Staender, A., & Morosoli, S. (2021). The sharing of disinformation in cross-national comparison: Analyzing patterns of resilience. Information, Communication & Society, 1-21.
34. Ivan, Cristina. (2013). "Resilience – The X Factor of the Organisational Endurance." In Intelligence in the Knowledge Society, Proceedings of the XVIIIth International Conference, by Irena Chiru Teodoru Stefan, 161-172. ANIMV Publishing House.
35. Ivan, Cristina, Irena Chiru, and Rubén Arcos. (2021). "A whole of society intelligence approach: critical reassessment of the tools and means used to counter information warfare in the digital age." Intelligence and National Security 495-511. doi:DOI: 10.1080/02684527.2021.1893072.

36. JamNews. (2017). Fake news in Moldova: fires, droughts, terror attacks and discredited politicians. septembrie 2017. https://jam-news.net/?p=59912.
37. Jeon, Youngseung, Bogoan Kim, Aiping Xiong, DONGWON LEE, and Kyungsik Han. (2021). "ChamberBreaker: Mitigating the Echo Chamber Effect and Supporting Information Hygiene through a Gamified Inoculation System." Proceedings of the ACIM on Human Computer INteraction. 1-26. doi:https://doi.org/10.1145/3479859.
38. Kratochvíl, Petr and Shakhanova, Gaziza (2020): The Patriotic Turn and Re-Building Russia's Historical Memory: Resisting the West, Leading the Post-Soviet East? Problems of Post-Communism, https://10.1080/10758216.2020.1757467
39. Lenin, V.V. (2018). "V. I. Lenin, Lessons of the Moscow Uprising." https://www.marxists.org/archive/lenin/.
40. Lewandowsky, Stephan, Ullrich K.H. Ecker, and John Cook. (2017). "Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era." Edited by Elsevier. Journal of Applied Research in Memory and Cognition.
41. Lewandowsky, Stephan, Ullrich K.H. Ecker, and John Cook. (2017). "Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era." Journal of Applied Research in Memory and Cognition. doi:Journal of Applied Research in Memory and Cognition.
42. Luxmoore, M. (2019) "Orange Plague": World War II and the Symbolic Politics of Pro-state Mobilization in Putin's Russia, Nationalities Papers, 47(5), https://doi.org/10.1017/nps.2018.48
43. Maftei, A., & Holman, A. C. (2022). Beliefs in conspiracy theories, intolerance of uncertainty, and moral disengagement during the coronavirus crisis. Ethics & Behavior, 32(1), 1-11.
44. Maschmeyer, Lennart (2021). Digital Disinformation: Evidence from Ukraine, CSS Analyses in Security Policy 278, 2021, 2 https://doi.org/10.3929/ethz-b-000463741
45. Martin, L. John. (2010). "Dinsinformation: an instrumentality in the propaganda arsenal." Political Communication 47-64.
46. McKay, Spencer, and Chris Tenove. (2020). "Disinformation as a Threat to Deliberative Democracy." Political Research Quarterly. doi:10.1177/1065912920938143.
47. Mediacritica, primul portal de educație mediatică. (2018). Moldova – teren fertil pentru fake news. iulie 11. Accessed iulie 30, 2018. http://mediacritica.md/ro/moldova-teren-fertil-pentru-fake-news/#prettyPhoto.
48. Monod, David. (2010). "'He is a cripple an' needs my love': Porgy and Bess as Cold War propaganda." Intelligence and National Security 1-14.
49. "Multiculturalism." Oxford Dictionaries. Accessed 08 5, 2014. http://www.oxforddictionaries.com/definition/english/multicultural.
50. Nabb Research Center Online Exhibits. n.d. "The Colors of Propaganda." https://libapps.salisbury.edu/nabb-online/exhibits/show/propaganda/what-is-propaganda-/the-colors-of-propaganda.
51. Nissembaum, Assaf, and Limor Shifman. (2015). "Internet memes as contested cultural capital: The case of 4chan's /b/ board." SagePub Journals 1-19. doi:DOI: 10.1177/1461444815609313.
52. Nye, J. S. (1990). Soft power. Foreign policy, (80), 153-171.
53. Paul, Richard, and Linda Elder. (2014). Critical Thinking: Tools for Taking Charge of Your Professional and Personal Life. New Jersey: Pearson Education.
54. Pavlik, John V. (2022). The Russian War in Ukraine and the Implications for the News Media, Athens Journal of Mass Media and Communications, 8: 1-17 https://doi.org/10.30958/ajmmc.X-Y-Z 1
55. Polygraph.info. (2018). "Polygraph." April 26. Accessed August 30, 2018. https://www.polygraph.info/a/fake-news-in-hungary/29194591.html.
56. Pressman, D. Elaine, and Cristina Ivan. (2019). Internet Use and Violent Extremism: A Cyber-VERA Risk Assessment Protocol. IGI Global.
57. Silverman, Craig and Kao, Jeff (2022). In the Ukraine Conflict, Fake Fact-Checks Are Being Used to Spread Disinformationhttps://www.propublica.org/article/in-the-ukraine-conflict-fake-fact-checks-are-being-used-to-spread-disinformation [Consulted: 17/01/2023]

58. Somers, Margaret R. n.d. The narrative constitution of identity:A relational and network approach. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/43649/11186_2004_Article_BF00992905.pdf?sequence=1.
59. Stănescu, Georgiana (2022). Ukraine conflict: the challenge of informational war, social sciences and education research review, 9 (1), 146-148 https://doi.org/10.5281/zenodo.6795674
60. Stoica, Cătălin Augustin, and Radu Umbres. (2020). "Suspicious minds in times of crisis: determinants of Romanians' beliefs in COVID-19 conspiracy theories." European Societies S246-S261. doi:https://doi.org/10.1080/14616696.2020.1823450.
61. TaskForce, EU East StratCom. (2020). Trends of the Week. Throwing Coronavirus disinfo at the wall to see what sticks . EU StratCom Task Force .
62. "The 2022 Code of Practice on Disinformation." European Commission . July 2. file:///C:/Users/User/Downloads/2022_Strengthened_Code_of_Practice_Disinformation_TeAETn7bUPXR57PU2FsTqU8rMA_87585.pdf.
63. n.d. Ukrainian tractor memes compilation. https://www.youtube.com/watch?v=hheLODstezM.
64. United Nations General Assembly. (2015). "Plan of Action to Prevent Violent Extremism, Report of the Secretary-General." A/70/674. Accessed September 20, 2020. https://www.un.org/en/ga/search/view_doc.asp?symbol=A/70/674.
65. University of Oxford. (2018). The Computational Propaganda Project. Algorythms, Automation and Digital Politics. http://comprop.oii.ox.ac.uk/.
66. Vázquez Liñan, Miguel (2010). History as a propaganda tool in Putin's Russia, Communist and Post-Communist Studies 43, 167–178.
67. Weisburd, Andrew, Clint Watts, and JM Berger. (2016). "Trolling for Trump: How Russia Is Trying to Destroy Our Democracy." War on the Rocks. https://warontherocks.com/2016/11/trolling-for-trump-how-russia-is-trying-to-destroy-our-democracy/.

1.3 Conspiracy Theories

Ruxandra Buluc, Cristina Arribas, Ana Ćuća

Abstract

The present section addresses one of the most pressing current challenges in fighting disinformation: conspiracy theories. Conspiracy theories have always existed in societies, however, at present, they have gained momentum due to their easy spread and appeal in social media. Moreover, they have begun to corrupt people's understanding of the world and their willingness to listen to experts and authorities in times of crisis and not only, thus threatening not only the further development of societies but also the very health and security of the communities they live in. The research looks into what conspiracy theories and their characteristics, what effects they have on societal progress and well-being when they become widely accepted. The limitations of the research are given by the fact that, as of yet, it is difficult to identify a rapid bulletproof method of countering their effects, and more work needs to be done in garnering trust in public institutions, authorities, scientists, in order for the public to view conspiracy theories as disreputable attempts to subvert the mechanisms of democratic societies. The response to conspiracy theories is firstly based on raising awareness to how widely spread they are, on what makes them attractive, and how people should respond to them, when they come into contact with them. Secondly, and in the long run, increasing the level of education of the population will lead to their ability to cope when high risk, little information crises occur, which will drain the fertile ground of uncertainty in which conspiracy theories bloom.
The result will be a step-by-step guide to conspiracy theories in order to enhance public comprehension of the phenomenon and raise awareness to the societal dangers it poses.

Main research questions addressed

- What is a conspiracy theory?
- What are the characteristics of conspiracy theories? Why are conspiracy theories attractive?
- How do conspiracy theories affect the common understanding of events?
- Why are conspiracy theories dangerous for societal cohesion?
- How can conspiracy theories be countered?

Definition of conspiracy theories

Understanding and countering the negative effects that conspiracy theories have on contemporary democratic societies means that first and foremost, it must become clearer what conspiracy theories are and how they can be distinguished from actual conspiracies that have and will continue to exist in society. Uscinski proposes a definition for a conspiracy: "a secret arrangement between two or more actors to usurp political or economic power, violate established rights, hoard vital secrets, or unlawfully alter government institutions to benefit themselves at the expense of the common good" (2018, 48). He also stresses the fact that a real conspiracy refers to events that proper authorities have determined that have actually occurred. The proper authorities have at their disposal the instruments needed to investigate and they are also comprised of people who have the verifiable and certifiable competencies and skills to evaluate and establish what events have actually happened. Problems arise in contemporary societies because there is an increasing distrust in competent authorities as well as in expert knowledge and this fuels the public suspicion of official explanations and their quest for alternative ones, which often contradict official reports and endorse conspiracy theories.

One very well-known example of a conspiracy theory is that 9/11 was an inside job. This conspiracy theory has multiple strands: a) 9/11 was planned by the American government; b) the American government knew in advance the attacks were going to happen and did nothing to prevent them; c) the attacks were, in fact, planned demolitions staged as terrorist attacks, in order to justify the invasion of Afghanistan and Iraq, and/or to curtail civil liberties by the measures that have been taken since, and/or to create a globalist government.

However, examples do not ease the difficulty of providing an accurate, synthetic and workable definition of conspiracy theories. One of the best known and most widely accepted is Uscinski's: a conspiracy theory refers "to an explanation of past, ongoing, or future events or circumstances that cites as a main causal factor a small group of powerful persons, the conspirators, acting in secret for their own benefit and against the common good" (Uscinski, 2018). Keeley approaches the definition of conspiracy theories from a logical point of view and as such characterizes them as unwarranted as they propose "an explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons – the conspirators – acting in secret" (1999, 116). Prooijen & van Lange (2014) and Douglas & Sutton (2018) also emphasise the secrecy and nefariousness of the plots expounded in conspiracy theories, as well as their explanatory prowess. Brotherton (2015) points out that conspiracy theories are "are easy ways of telling complicated stories" which provide a means of eliminating complexity and clearly identifying causal relationships and perpetrators.

As previously mentioned, conspiracy theories go against official explanations provided by legitimate, epistemic authorities (Brotherton & Eser, 2015). They make use of weak evidence, small unaccounted for details, endow the conspirators with sinister goals and above average competence. However, no matter how outlandish they may appear, they have serious social consequences: reduced civic engagement, negative attitudes towards environmentalism, vaccination etc.

Moreover, as Cassam (2021) and Oliver & Wood (2014) argue, conspiracy theories have political motivations and promote political ideologies, by providing the compelling explanatory narratives that sway public conviction in the desired ideological direction.

In brief, we propose the following definition for conspiracy theories: they are explanatory causal-based, ideologically laden narratives which depict significant social events or crises as perpetrated by a group of powerful secret actors who solely follow their own nefarious interests, irrespective of the good of the masses.

Characteristics of conspiracy theories:

1. Conspiracy theories are *speculative*, meaning that they are "based on conjecture rather than knowledge, educated (or not so educated) guesswork rather than solid evidence" (Cassam, 2021). This aspect is doubled by other characteristics stemming from the fact that they are based on fringe science: they are esoteric, as in they promote strange alternative explanations to official stories. They rely on circumstantial rather than direct evidence, on conjecture rather than solid evidence.
2. Conspiracy theories are *contrarian* by nature (Cassam 2021, Brotherton, 2015, Wood & Douglas, 2013, Oliver & Wood, 2014, Keeley, 1999). They run counter to the official narrative or view, to the obvious, plausible and acceptable explanations of events. The obvious answer is never correct, as conspiracy theories cast doubt on everything, even the best scientifically supported explanations. An example is this direction is the flat earth conspiracy theory which claims that the scientifically proven fact that the Earth is round is a conspiracy. Instead, conspiracy theories identify the source of any event or of scientific facts in unseen, malevolent forces who aim to harm people and societies and hide their nefarious actions.
3. As a consequence of the fact that they are based on pseudo-science and fringe science or ignorance of science, conspiracy theories are *amateurish*, as Cassam (2021) explains, referring to the qualifications of amateur sleuths and internet detectives who produce and promote them.
4. Conspiracy theories are *premodern* (Cassam, 2021, Keeley, 1999, Douglas and Sutton, 2018, Oliver & Wood, 2014), meaning that they attempt to impose order in a random, complex, uncontrollable world in which events, crises are seen by conspiracy theorists to occur as a result of evil machinations not as a result of a conjunction of numerous factors, causes and even coincidence that cannot be and are not controlled by any one person or group of persons. They are based on a Manichean, simplistic worldview, clearly divided into good and bad forces, with no grey areas, and no place for randomness. The lack of control that people experience when faced with tragic events is compensated for by attributing agency, be it malevolent, to a small group of powerful people who could bring about doom. Conspiracy theorists are not paranoid or delusional, but they do experience the need to identify and/or assign intentionality in the environment. They need to rely on the idea that things happen for a reason, that there is a design behind the randomness of events, that a pattern can be identified in haphazard stimuli.
5. Conspiracy theories are *self-sealing* and *self-sustaining* belief bubbles which makes them unfalsifiable (Cassam, 2021, Brotherton, 2015, Vermeule & Sunstein, 2009). They are difficult to challenge because any counterargument is met with the challenge "They would say that, wouldn't they?" which basically incorporates the contrary information into the conspiracy theory itself. This type of logic is unassailable, as any contrary evidence is interpreted as proof that the conspiracy is at work, hiding its machinations from the eyes of the public with a smokescreen of counterarguments. "Self-insulated logic which makes them immune to refutation and they actually thrive on it" (Brotherton, 2015).

Being self-sealing and self-insulating also makes conspiracy theories strong since they are able to incorporate any apparently anomalous piece of information into the unifying theory they propose. As Keeley (1999: 118) explains, conspiracy theories operate with unaccounted for data (data which is not included in the official explanation of the event) and contradictory data (data which goes against the official explanation of the event). These two types of data give rise to questions, which can, in turn, lead to conspiracy theories. As Brotherton (2015) explains, in essence, conspiracy theories are unanswered questions, which try to reveal hidden plots and to alert the masses that the truth is not the one officially presented, but different, always somewhat out of reach, just beyond the next data incongruence.

6. For these reasons, conspiracy theories are very *nuanced* and *complex*. The simplest explanation is never sufficient because it cannot account for everything, it cannot account for randomness and coincidence and it does not provide the all-encompassing explanation that the conspiracists' premodern mindsets seek. "Unified explanation is the sine qua non of conspiracy theories. Conspiracy theories always explain more than competing theories, because by invoking a conspiracy, they can explain both the data of the received account and the errant data that the received theory fails to explain" (Keeley, 1999, 119). The rule of logic Occam's razor is suspended, due to the fact that simple does not mean fully explanatory and, therefore, more complexity is needed, even if it is not warranted.
7. In order to reach its end, a conspiracy is, by definition, *unknowable* to and *untraceable* by the larger public. This leads to the contradictory nature of conspiracy theories and theorists, who, on the one hand, view conspirators as all-powerful masterminds who are able to protect secrets, control the population, are responsible for all the bad things that happen in the world, etc., and, on the other hand, the conspiracy theorists overvalue their own abilities to catch them, to divine their plans and intentions. (Cassam, 2021, Brotherton, 2015, Vermeule & Sunstein, 2009) This raises the question "If the conspirators are so clever, how come they have been rumbled by a bunch of amateurs?" (Cassam, 2021) This question remains unanswered and conspiracy theorists are unfazed by it as they believe they are engaged in a David vs Goliath struggle and that the apparently weak, but in fact vigilant person can outfox the greatest and most potent conspirators. Grimes (2016) points out that it is human nature for conspirators to leak information in the case of real conspiracies. Secrets are hard to keep due to human nature, but once the flaws in human nature are also doubled by technological weaknesses which allow for leaks or hacking, secrets become increasingly hard to handle. Moreover, the more time passes, the more likely people are to talk more freely about that secret, which is why, Grimes argues, it is not feasible to believe in long-standing conspiracy theories. If they had truly existed, they would have become evident.
8. Conspiracy theories form a *monological belief system* (Goertzel, 1994; Wood, Douglas & Sutton, 2012; Prooijen & van Lange, 2014). This means that each belief supports every other belief, and the more conspiracies a monological thinker believes, the more likely they are to believe new ones as well, regardless of their topic. As Brotherton (2015) further

explains, the conspiracist mindset operates according to the slippery slope logic: if one conspiracy theory is true, it could become evidence for others being true. Wood, Douglas & Sutton (2012) and Douglas & Sutton (2018) have discovered the reason behind this. More precisely, the researchers discovered that this monological belief system is not determined by individual conspiracy theories, but by "agreement between individual theories and higher-order beliefs about the world" (Wood, Douglas & Sutton, 2012, 768), such as the idea that the authorities are deceptive and act against public good. Therefore, if a new conspiracy theory is presented in which authorities are seen as being manipulative and secretive it is more easily accepted if the recipients already hold this belief, and, in this case, it will not matter if it contradicts another previously held conspiracy theory.

9. Conspiracy theories purport that people are not merely kept in the dark, they are being *actively fooled* by the authorities, as all appearances are misleading, and the elites do not have the people's best interests at heart. Official accounts are only meant to distract public attention from what powerful elites have actually planned, and their intentions are invariably evil and nefarious (Brotherton, 2015; Oliver & Wood, 2014)
10. Conspiracy theories are culturally based. This refers not only to the cultural knowledge and customs derived from the evolution of a specific society, but also to the cultural values that shape the organization and functioning of that particular society. In a broad cross-cultural study, Adam-Troian et al (2020) examined how culture may influence belief in CTs employing culture-as-situated cognition theory, which investigates the ways in which individual cognitions are activated by the particular cultural context the individuals find themselves in. They measured six values: power distance, individualism, masculinity, uncertainty avoidance, long-term orientation and indulgence and they concluded that cultural values have a unique predictive power on conspiracy theory belief. More precisely, masculinity and collectivism were robust positive predictors of conspiratorial belief across countries, operationalizations, and levels of measurement.

    Moreover, as Cassam (2021) has also indicated, conspiracy theories have an important ideological and political component. Radnitz & Underwood (2015) determined that political values guide belief formation and can anchor and even limit a search for information to that information that is consistent with the individual's already held values. Therefore, liberals believe different conspiracy theories than conservatives because they often hold divergent values regarding societal political organization.

One of the first works examining the psychological determinants of belief in conspiracy theories was published by Hofstede and he stated that paranoia was mainly responsible for this belief. Since then, numerous studies have been carried out to understand what personality traits foster adherence to conspiracy theories and this research has demonstrated that the situation is more complex and that people who do not exhibit any divergent personality traits might also adhere to conspiracy theories. Therefore, a more nuanced understanding of the personality traits that foster it is required. In fact, recent studies have revealed that it is a combination of psychological and cognitive mechanisms that are in fact facilitators and aggravating factors for conspiracy theorists.

Psychological factors conducive to belief in conspiracy theories

Conspiracism ideation can be associated with *paranoia*. As the experiments performed by Brotherton & Eser (2015) and Darwin et al (2011) Swami et al. (2011, 2013) and Goertzel (1994) prove people with higher scores on the paranoid ideation scale are more vigilant by nature, as they constantly look for signs of hostility directed against them and might misinterpret innocent coincidences or sequences of events as conspiracies, and they are also less likely to accept official explanations for those events. However, mild paranoia is associated with other psychological traits: low self-efficacy (Brotherton & Eser 2015), low self-esteem (Brotherton & Eser 2015, van Prooijen & van Lange, 2014), dissatisfaction with life (Brotherton & Eser 2015), higher levels of anxiety (Vermeule & Sunstein 2009, Radnitz & Underwood, 2015, Brotherton & Eser, 2015), distrust of others (Goertzel, 1994), insecurity about elements of the environment (Goertzel, 1994, Moulding et al, 2016).

Conspiracy theorists also experience powerlessness in the face of random events that affect their lives to varying degrees (Brotherton & Eser, 2015). This leads to people needing to find explanations, even based on attribution errors, which place the blame and foster hostility to the outgroups, to those who are different, in any significant way, from the conspiracy theorist. Goertzel (1994) also found that anomia (defective moral sense) is associated with belief in conspiracy theories, because it reflects the feelings of alienation and disaffection with the system that conspiracy theorists experience. When resorting to conspiratorial explanations, people focus the blame on a tangible enemy, and thus the problems become less abstract and impersonal. Moulding et al (2016) further explained the relation between anomia and conspiracy theories by positing that conspiracists observe that moral standards are not fixed, they are ever-changing according to the temporal and situational considerations, therefore reaching the conclusion that the world is a bad place that actually conspires to hurt them.

This is doubled by the fact that the human mind tends to look for intentional causation based on perceived benefits for certain parties (Vermeule & Sunstein, 2009, Radnitz & Underwood, 2015), which conspiracy theories so readily provide in the guise of the villainous conspirators whose aim is to destroy the lives of the unaware masses to serve their own interests. There is a pervasive human tendency to think that effects are caused by intentional action, especially by those who stand to benefit (the "cui bono?" maxim), and for this reason conspiracy theories have considerable but unwarranted appeal. (Vermeule & Sunstein 2009, Radnitz & Underwood, 2015, Oliver & Wood, 2014). Brotherton (2015) explains that this search for the culprit in complex or random situations is an expression of the need for control and order, as it is easier for people to accept that there is an intention behind those events, even if it is malevolent, than to accept randomness. This is an instance of compensatory control. Moreover, conspiracy theories also serve as readily available and easily understandable answers to complex issues, and provide a scapegoat or blame target, but not haphazardly. They reflect the ideological foundation that the supporters of the respective conspiracy theories share, and they divide the world into simple dichotomies such as us vs them, irrespective of who they might be in that particular instance (migrants, political parties, foreign governments, multinational corporations, etc.). These feelings of being oppressed, or excluded, or duped create strong community bonds among conspiracy theorists, as the victimized group.

Douglas & Sutton (2018) further refine the search of psychological traits that foster conspiracism as they focus on one particular aspect of social anxiety, namely anxious attachment, which is correlated with a preoccupation for security, negative views of outsiders, sensitivity to threats and overestimations of threats. Moreover, they also identify the need for uniqueness (the need or desire to be different from other people) as a predictor of belief in conspiracy theories.

Cognitive factors conducive to belief in conspiracy theories

The human brain does not handle not knowing or not understanding well, even in complex and incomprehensible situations. It actively searches for an explanation or for a solution so that at least apparent control over the situation is exerted. Brotherton (2015) explains that this is due to a "metacognitive glitch", that drives the brain to fill in any blanks with familiarity and beliefs rather than actual knowledge, in the guise of better or poorer guesses if nothing else is available. In order to do so the brain operates with several cognitive mechanisms that could not only provide understanding quickly, but could also make people more prone to believing conspiracy theories. Among these Cassam (2021) identified the intentionality bias which refers to the tendency to assume things happen because they were intended rather than accidental; the confirmation bias which entails the tendency to look for evidence that supports what one already believes while ignoring contrary evidence; and the proportionality bias which presupposes the tendency to assume that the scale of an event's cause must match the scale of the event as such (also supported by research performed by Brotherton & French 2014, Kahneman & Tversky 1972).

These mechanisms aid the formation of connections and the identification of patterns without conscious control. Causal relationships are established in light of what people believe rather than what they know or the facts available to them, because the human brain has been evolutionary trained to notice coincidence but try to infer a cause, based on what is already available to it. And beliefs are an integral part of human cognition and they are more stable and more widely applicable than facts. Therefore, they are an available source for explanations, no matter how spurious they may be for a certain situation. This cognitive mechanism is known as the conjunction fallacy (Tversky & Kahneman 1983) and it is an error of probabilistic reasoning which leads people to overestimate the likelihood of events occurring together. Brotherton & French 2014 conducted research to evaluate the extent to which the conjunction fallacy is related to conspiracism and discovered that people who believe in conspiracy theories committed more conjunction violations (spurious assignation of causation) than people who were less prone to conspiracist ideation.

Moreover, Brotherton (2015), Douglas & Sutton (2018), Douglas & Sutton (2011), explain that the brain employs the mechanisms of intention detector and projection. Projection refers to the fact that people try to understand what others are thinking and/or doing not solely based on the latter's actions, but on what they would do if they were in that situation, on putting themselves into other people's shoes. However, this brings about the projection, sometimes faulty, of one's own beliefs and emotions on other people who might actually not relate to them. The distorted lens of projection may lead to an understanding of a situation that creates a false consensus, does not allow for differences and randomness and may, in certain situations, reinforce the us vs them dichotomy. As Douglas & Sutton (2011, 547) explain "these results revealed that personal willingness to engage in the conspiracies predicted endorsement of conspiracy theories. Machiavellianism also predicted endorsement of conspiracy theories. Finally, the relationship between Machiavellianism and conspiracy beliefs was fully mediated by participants' willingness to engage in the conspiracies themselves."

Pytlik et al (2020) also identified the cognitive mechanism of jumping to conclusions (a fast, heuristic thinking style) as a predictor for conspiracism, more precisely for people's tendency to believe that conspiracies are the roots of important society altering events. Jumping to conclusions is also associated with trust in one's intuition, which leads to accepting "simple, yet satisfying narratives", involving a small group of individuals who pull the strings in society, which conspiracy theories provide. Analytical examination of the facts to ponder upon the ways in which various entities interact to create any given situation is time consuming and effortful, which is why jumping to conclusions, with the readily available explanations that result from this type of thinking, is more easily acceptable than admitting that one might not know or understand everything. Radnitz & Underwood (2015) came to the similar conclusion, that when faced with uncertainty and ambiguity, people make "snap judgements" with respect to assigning trust.

Prooijen & Jostmann (2013) examined the explanatory function of conspiracy theories. They reflect a systematic method of information processing, which forms a clear connection between evil conspiracies and threatening events, thus turning into sense-making processes that help people manage the uncertainty and the immorality they perceive exists in the world. "Uncertainty makes people more attentive to the morality of the actions of authorities when making sense of a threat to the social order. As such, uncertainty increases the extent to which people make inferences about the plausibility and the implausibility of conspiracy theories based on the morality of authorities' actions" (Prooijen & Jostmann 2013, 110). The world thus returns to being predictable, orderly, comprehensible, that is reflective of their monological belief system (see conspiracy theories characteristics).

Clarke (2002) explains that it is difficult for people to relinquish conspiracy theories because of "a fundamental attribution error" that they commit. He proposes that the reason why conspiracy theories are more attractive explanations than non-conspiratorial ones lies in the distinction between dispositional explanations of behavior – based on personality features, and situational ones – based on the characteristics of a situation, and on the fact that, more often than not, people overestimate the importance of the former to the detriment of the latter. More precisely, dispositional explanations are perceived as being more in tune with how people think and react, with their intentions, and also exhibiting more temporal continuity. "Dispositional explanations can relate the occurrence of events within the context of an intended plan" (Clarke 2002 146) usually based on an attribution error with respect to various individuals' intentions. Meanwhile, situational explanations only appear to function in a limited context, they usually form the bases of official accounts of a situation and can lack unificatory power, as they cannot be projected beyond those coordinates.

Main challenges in countering conspiracy theories

The need to counter conspiracy theories stems from the fact that their effects on society are grave. They challenge truth, and consensual truth matters greatly in a society as it is the foundation on which constructive and progressive dialogue is built. Without such dialogue, democratic societies at least are thwarted in their development by the polarization of the citizens who find themselves unable or unwilling to interact with others who have diverging opinions. If there is no common denominator of understanding and no common reference points, then debate becomes impossible, and arguments deteriorate into quarrels. All aspects of societal knowledge and function can be affected by conspiracy theories: science is altered when people believe that scientists are actually corrupted representatives of big corporations, the democratic processes suffer when people exercise their voting rights based on conspiracy theories and not facts and data, society is harmed when policies are enacted not based on knowledge but on conspirational beliefs, international relations suffer when disinformation outweighs facts and real events.

As dangerous as they are, the main challenges regarding conspiracy theories stem from the fact that they are difficult to counter. Despite the fact that their consequences for society are perilous indeed, they also appear very resistant to being debunked. Researchers into the field are unanimous in their assessment that effective debunking strategies of conspiracy theories should be multifaceted and include both a political and an intellectual dimension (Cassam 2021) to which others argue that an emotional component should also be attached as conspiracy theories reflect identity forming beliefs and, therefore, supporters are likely to feel aggrieved when facing counterarguments.

1. The *intellectual* dimension of a debunking strategy should focus on constantly rebutting the theories, by telling the truth. The truth may not dissuade die-hard conspiracy theorists, but may make it clear for the undecided that the conspiracy theory does not actually account for the events as they took place. West (2018) explains that rebuttal should be reinforced by the constant reference to the trustworthiness (or lack thereof) of the sources that conspiracy theorists gather their information from. If the source can be shown to be wrong on any account, then they might begin to question its reliability on all accounts. Moreover, conspiracy theorists should be exposed to new, accurate information constantly, so as to challenge their beliefs and possibly make them reassess them.
2. The *political* dimension of the debunking strategy should focus on exposure of any political interests the conspiracy theory might be serving, thus proving that it is part of political propaganda and not the truth. The ideological component of the conspiracy theory should be revealed and criticized, and people should be made aware of the fact that the respective theories are merely a political tool for a certain interested party to attain a benefit. Vermeule & Sustein (2009) explain that in order not to trigger a backfire effect and make a conspiracy theory even more popular during attempts to debunk it, authorities should not focus on debunking one particular such theory, but rather an ensemble of such theories, more precisely their points of commonality. Moreover, education with respect to debunking conspiracy theories and any form of disinformation should start as early as possible so as

to prepare future citizens with the instruments they need to accurately assess the information they are presented with, and separate facts from lies and misconceptions.

3. West (2018) provides a *personal* interaction dimension to the strategy of debunking conspiracy theories. He proposes three steps that could be undertaken to this end:
   a) Maintain effective dialogue which means that the debunker needs to understand what the conspiracy theorists are thinking and why, to be polite, respectful, open, to attempt to find common ground so as to validate their concerns if not their manifestations. Aggressive behavior will sever all lines of communication and have the backfire effect of actually strengthening conspiracy theorists' views.
   b) Supply useful information which could counter the backfire effect, by showing the conspiracy theorists what mistakes they have made, why their sources may not be reliable, what information about the topic they missed, and what other details on the topic are available, thus helping them gain perspective.
   c) Give it time means that the change cannot and does not take place immediately, and that patience and reiteration are required.

Of these three stages, arguably the most important is to build back common ground. As previously mentioned, the greatest challenge with conspiracy theories is that they erode the common ground vitally important for communication and progress in a society. A polarized society cannot reach consensus on anything, as dialogue is impossible with no common framework of understanding of how the world functions. Dennett (2014) offers a three-step process to enable the rebuilding of common ground:

a) Re-express the conspiracy theorists' position better than they do themselves, based on the principle of charity. This means that by restating the argument even better than initially presented, the debunker proves understanding, does the work to make the conspiracist details actually work, so that when the flaws are revealed, the conspiracy theorists are more likely to listen to them because they come from a person who understood them and what they were saying, that had built common ground.
b) List points of agreement, especially uncommon points, through a gradual exploratory process, that will slowly and patiently take the debunker through the arguments, until such commonalities are identified. They could be specific or general, but they are almost always there, and they once more set a stable common ground from which to start. If in the respective conspiracy theory none such points could be identified, then another more uncontroversial topic could be explored so as to have the needed starting point of agreement.
c) Mention anything that you have learned from the conspiracy theorists as this increases rapport, proves that real communication has taken place, and thus common ground is reinforced. This step might also include a validation of the conspiracy theorists' genuine concerns so that they feel heard and understood, rather than high-handedly dismissed.

Vermeule & Sustein (2009) also suggest a more radical and somewhat difficult to implement tactic for breaking up the hard core of extremists who supply conspiracy theories: "cognitive infiltration of extremist groups, whereby government agents or their allies (acting either virtually or in real space, and either openly or anonymously) will undermine the crippled epistemology of believers by planting doubts about the theories and stylized facts that circulate within such groups, thereby introducing beneficial cognitive diversity" (Vermeule & Sunstein, 2009, 219). However, this would be a very dangerous tactic to apply, because if the infiltrated agent were to be uncovered, then the group would take it as further proof that there is a governmental conspiracy at work, which would radicalize their belief in the conspiracy theory even further.

Debunking conspiracy theories may be difficult and very time-consuming, however, it is more needed than ever, as people seem more likely than ever to hide in their respective bubbles and break all forms of communication with anyone who disagrees with them. Such polarization, not solely along political lines, but also along understanding of facts and relation to reality, can only lead to dysfunctional societies.

1.3.1 Case study (1) The Great Dacians whose history is denied

Historical records show that the Dacians lived in the space currently occupied by Romania. The Roman policy was to quell any possible riots on the edges of its empire, and the Dacians' resistance to Roman conquest is documented. Moreover, the Romans had heard rumors of the Dacian gold, and they needed to restore their financial resources. Consequently, emperor Traian decided to attack Dacia. The Roman legions arrived south of the Danube in 101 AD and crossed the river in three places. The Dacian ruler, Decebal, was waiting for them at Tapae. After an initial victory, the Dacians are defeated and they become Roman subjects under Decebal's rule. However, as Decebal does not respect the terms of the peace treaty and attempts mutiny, the Roman legions return in 105 AD and conquer all of the main Dacian strongholds, including Sarmisegetusa, the capital. The Dacians are defeated and Decebal kills himself. The Romans conquer a part of Dacia, of the southern and western territories. The Romans remained in Dacia until 271 AD.

An example of a conspiracy theory that is specifically Romanian can be summarized as "Romania is the birthplace of Europe"[9], as the Dacians are the ancestors of most civilizations on earth. The Dacians are the ones who migrated west and eventually formed the Roman Empire, as well as east, reaching as far as Japan and India. Moreover, they are the inventors of writing (the Tartaria tablets are the first written records), the wheel, the plough, the cart, mining equipment, among others. The Thracians and the Dacians (a subgroup of the former) represent the oldest and highest culture on earth, precursor of the Sumerian civilization, and also the most numerous (180-200 tribes), spread all over Europe, Asia and Africa[10].

The proponents of this theory base their claims on a) alleged linguistic incongruities and b) depictions of Dacians in Roman sculptures.

[9] https://lupuldacicblogg.wordpress.com/2017/06/06/istoria-adevarata-a-romaniei-dacii-sunt-primii-oameni-care-au-populat-europa-dacia-nu-a-fost-cucerita-niciodata-in-intregime-stramosii-romanilor-sunt-dacii-nu-romanii/

[10] Idem

As far as linguistic incongruities are concerned, they are meant to explain away one of the main counterarguments against this conspiracy theory, namely that Romanian is decidedly a Latin language. The problem conspiracy theorists have with the official historical narrative is that Dacia was not entirely conquered, and it was under occupation for a relatively short period of time (approximately 170 years). Therefore, it is not possible that almost the entire Dacian language was lost, the Dacians having learnt vulgar Latin to interact with the conquerors, and only sporadic words remain. The conspiracy theorists' claim is that actually Dacian was an older version of Latin and for this reason, when the Romans arrived in Dacia, they did not have to assimilate the Dacian population and their language as they were already quite similar. What the conspiracy theorists do not take into account is that the Dacians needed to communicate with the Romans, thus they learnt their language. They also intermingled and spread the new language to other tribes as well, as a means of boosting commerce.

Secondly, the depictions of Dacians in Roman post-conquest sculptures in Rome, show them as standing tall and proud and still wearing their specific headgear, which, if they had been truly conquered, they would have been forced to remove as a symbol of their humiliation. However, the fact that they were allowed to keep it, as the statues prove, means that in fact the Romans did not conquer Dacia, but rather forged an alliance with them, in which they recognized the Dacians as their forefathers and their merits in the formation of the Roman Empire.

This conspiracy theory is an example of nationalist photochromism and has no foundation in historical documents. It is also fueled by Romanian exceptionalism and by a part of the population's belief that foreign malevolent forces are constantly attempting to keep Romanian superiority, ingenuity, greatness hidden and strive to harm Romanian development so as not to affect their own interests. This conspiracy theory views Romanians as both exceptionally gifted as well as eternal victims of plots to undermine their greatness and their development. Victimization and blame assignation are the fuel that drive this conspiracy theory.

1.3.2 Case study (2) The murder of Daphne Caruana Galizia

Daphne Caruana Galizia was a very well-known Maltese reporter, editor, columnist and blogger. Her blog Running Commentary had a very high reach, comparable to the main media houses in Malta. Her continuous challenging of political power structures through her reporting on corruption, sleaze and crime, made her both liked and disliked by many.[11] Throughout her career, Daphne Caruana Galizia received threats and was the target of several forms of harassment because of her journalism. On 16 October 2017 Daphne was assassinated by the triggering of an explosive device planted under her car seat outside her home in Bidnija, Malta. The investigation of her

[11] Borg, J. (2017). Daphne Caruana Galizia obituary. *The Guardian.* Available at: https://www.theguardian.com/media/2017/nov/21/daphne-caruana-galizia-obituary

assassination further exposed the corruption of the government and institutions who were accused in a public inquiry of having created an atmosphere of impunity[12].

A Maltese businessman, Yorgen Fenech was charged with having been the mastermind behind her assassination, but the trial is still ongoing. Three other people, Alfred and George Degiorgio and Vince Muscat were convicted of making, planting and detonating the car bomb that killed the journalist. In spite of the fact that the Police Commissioner has declared that all suspects in the case have been arrested and many of them have already been convicted, the case is still causing many controversies.

One such controversy is the conspiracy theory developed by Simon Mercieca, an Associate Professor at the University of Malta who employs his blog *Simon Mercieca's FreePress* to share a number of fake news and conspiracy theory on a wide variety of subjects, ranging from Maltese politics to COVID-19. According to him, Yorgen Fenech is innocent, while Daphne's assassination was organised by her husband (Peter Caruana Galizia) and her son (Matthew Caruana Galizia). According to his theory, the Caruana Galizia family is "hampering the investigation process and the court's operations so that the whole truth behind Daphne Caruana Galizia's murder will never be known. To achieve this scope, main witnesses of the prosecution, including Matthew Caruana Galizia are using the media and giving interviews to siphon issues to fit their agenda and condition the public."[13]

Mercieca's theory has all the usual characteristics of conspiracy theories. Firstly, it is speculative. There is no actual evidence that the Caruana Galizia family is hampering the court's operations, nor that they have in any way deceived the prosecution or the public. Mercieca claims that Daphne's son, Matthew, decided to take the law into his hands and destroy potential key evidence, albeit there was never any official information to support this claim. The key word in this theory is potential**.** Mercieca's theory relies on conclusions which are drawn based on circumstantial evidence, offering an explanation that is different from the official media reports and from the evidence presented in court. This leads to the second conspiracy theory characteristic – it is contrarian. Mercieca capitalises on the fact that the Maltese public is still divided on the subject of Daphne's assassination, with some groups arguing that the investigation and prosecution have not been carried out in the most transparent and efficient manner. However, instead of aligning himself with those who sought justice for the journalist's assassination and her family, his conspiracy theory argues the opposite that while justice has not been served the victim has been the Maltese businessman accused of murdering Daphne Caruana Galizia, namely Yorgen Fenech. He claims that Daphne's family have intentionally hindered the investigation and sought to gain money from the investigation, by putting the blame on a well-known Maltese businessman.

[12] Camilleri. N, Schembri Orland K. (2021). Public inquiry holds The State responsible for Daphne Caruana Galizia's murder. *Malta Independent*. Available at ttps://www.independent.com.mt/articles/2021-07-29/local-news/Public-Inquiry-holds-state-responsible-for-Caruana-Galizia-s-death-6736235558.

[13] Mercieca, S. (2021). The Caruana Galizia family should be made responsible for the money that is being wasted by the state because of their antics. *Simonmerecia.com.* Available at: https://simonmercieca.com/2021/11/25/the-caruana-galizia-family-should-be-made-responsible-for-the-money-that-is-being-wasted-by-the-state-because-of-their-antics/

Moreover, he argues that Daphne’s family didn’t put pressure on the authorities to bring Yorgen Fenech to justice, in the hope that as more time passes they will be able to build the case on false information and hide traces that could lead back to them.[14] This argument goes against official information and ignores existence evidence gathered in the case and presented during the criminal trial, including the testimonies of the people convicted for making, planting and detonating the car bomb that killed the journalist. Instead, the theory develops a scenario of demonization, whereby the real “malevolent forces” involved in the case are Daphne’s family, who not only harmed Daphne but are now harming Yorgen Fenech and the Maltese society in general.

This conspiracy theory shows how, by taking a complex situation, one may very easily build a theory that will be widely spread on the premise of a simplistic view assigning intentionality to Daphne’s family to not only harm her but the whole Maltese society. This conspiracy theory is built around self-sealing conclusions, built on information taken out of context, which makes it hard to refute. Moreover, the conspiracy theories are built one upon another, as in Mercieca uses the idea of “malevolent forces” seeking to discredit him (e.g. they would say that, wouldn’t they?) as an argument against all criticism received in relation to the other ideas promoted. This can also be seen as an indicator that he does not have any other counter-arguments/evidence that he can bring in support of his theories.

1.3.3 Case Study (3) Conspiracy theories on the August 17 terrorist attacks

On August 17, 2017, the worst terrorist attacks in Spain, since the Madrid train bombings of March 2004, took place in the Catalonian towns of Barcelona and Cambrils, with 16 deaths and more that 120 wounded.[15] The attacks temporarily coincided with preparations for the 1 October illegal referendum conducted by Catalonian secessionist parties and the Catalonian local administration.

In this context, some national and local news media, together with pro-independence political actors introduced the idea of a covert participation of Spanish intelligence in the attacks. The conspiracy theory was compounded by the decision made by the highest Spanish court (Audiencia Nacional) that rejected the request made by one of the victims' lawyers (and pro-secessionist member of the Catalonian parliament) to investigate the alleged connections. These two events led to the creation of the basis of an alternative explanatory theory.

This conspiracy theory has had different versions but in essence all of its iterations attribute the responsibility of the attack to the Spanish state through its intelligence services. Sometimes, alternative theories point out to direct implication, but other versions of this conspiracy theory authorities and security services are accused of negligence and lack of action when counting with

[14] Mercieca, S. (2021). The truth is out: Yorgen Fenech did not want Daphne Caruana Galizia killed. *Simonmerecia.com.* Available at: https://simonmercieca.com/2021/11/26/the-truth-is-out-yorgen-fenech-did-not-want-daphne-caruana-galizia-killed/

[15] https://www.rtve.es/noticias/20170827/atentados-barcelona-cambrils-dejan-14-muertos-mas-120-heridos/1599361.shtml

intelligence on the impending attacks. As far-fetched as it may appear, this malicious narrative can be captured in the following sentence:
"The sewer of the state work to harm Catalonia"[16]

It is easy to note the similarity of this conspiracy theory to those related to the "Deep state" that have circulated in other countries.

Description of the facts

The events began on August 17, 2017 on the Paseo de Las Ramblas in Barcelona where at 5:00 p.m. a van ran into a crowd of passers-by. On board was a single driver who managed to flee. Hours later, the Islamic State claimed responsibility for the attack through the Amaq news agency. During the early morning of August 18, in the nearby town of Cambrils (Tarragona), another vehicle broke into the promenade, and ran over five pedestrians and a policeman. The vehicle was intercepted, and the terrorists shot dead[17].

These events were connected to the explosion of the previous day (August 16) in a house of Alcanar (Tarragona, Spain). Two people died because of those explosions, including Abdelbaki Es Satty, leader of the cell and imam of Ripoll as it was later discovered. Another terrorist who was later tried for the attacks was wounded[18]. According to the instructions carried out and the content of the judicial sentence, a large attack with "van bombs" was being prepared in the Alcanar house. The explosion precipitated the subsequent attack in the Ramblas, and Cambrils since there was an ongoing police investigation and perpetrators knew they might get arrested. [19]

Origin of the conspiracy theory

On 16 July 2019, almost two years after the attacks, a well-known Spanish digital newspaper published the results of a journalistic investigation reporting alleged evidence on the fact that the terrorist cell was being subject to surveillance by Spanish intelligence and that the Imam of Ripoll was a human source for the Spanish National Intelligence Centre (CNI).[20] The journalistic pieces included images of a hypothetical surveillance report prepared by the Spanish intelligence, as well as the messages allegedly exchanged between the imam and the service through the dead drop system.[21]

Elements exploited to manufacture the alt version

Conspiracy beliefs have also been linked to the need for cognitive closure (Marchlewska, Cichocka and Kossowska, 2018; Leman and Cinnirella, 2013), especially when events lack a clear official explanation (cited in Douglas et al., 2019, 7). This is valid for this case study analysis in which the need for knowledge and clarifications on the terrorist attacks acts as a catalyst for the

[16] https://english.elpais.com/elpais/2012/11/21/inenglish/1353502495_401040.html

[17] https://www.elmundo.es/internacional/2017/08/17/5995eecb468aeb39228b45cd.html

[18] Ibidem

[19] https://www.theguardian.com/world/2021/may/27/three-men-jailed-over-2017-catalonia-terror-attacks

[20] See: https://www.publico.es/politica/exclusiva-iman-ripoll-1-cerebro-masacre-ramblas-confidente-cni-dia-atentado.html

[21] El independentismo abraza la teoría conspirativa sobre el 17-A, El País: https://elpais.com/ccaa/2019/07/26/catalunya/1564165566_085516.html?event_log=oklogin

spread of this conspiracy theory. In the absence of all pieces of information – something not unusual in criminal and intelligence research – different actors, sometimes with political interests, fill the gaps with unsupported assumptions.

References:

1. Adam-Troian, J., Wagner-Egger, P., Motyl, M., Arciszewski, T., Imhoff, R., Zimmer, F., ... & van Prooijen, J. W. (2021). Investigating the links between cultural values and belief in conspiracy theories: The key roles of collectivism and masculinity. *Political Psychology*, *42*(4), 597-618.Darwin, H., Neave, N., &
2. Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in psycho logy*.
3. Brotherton, R., & French, C. C. (2014). Belief in conspiracy theories and susceptibility to the conjunction fallacy. *Applied Cognitive Psychology*, *28*(2), 238-248.
4. Brotherton, R. (2015) *Suspicious Minds: Why We Believe Conspiracy Theories*. Bloomsbury Sigma.
5. Brotherton, R., & Eser, S. (2015). Bored to fears: Boredom proneness, paranoia, and conspiracy theories. *Personality and Individual Differences*, *80*, 1-5.
6. Butter, M., & Knight, P. (Eds.). (2020). *Routledge handbook of conspiracy theories*. Routledge.
7. Cassam, Q. (2021). Bullshit, Post-truth, and Propaganda. *Edenberg, E. and Hannon, M., Political Epistemology*, 49-63.
8. Cassam, Q. (2021). Conspiracy Theories. Polity Press.
9. Marchlewska, M., Cichocka, A., & Kossowska, M. (2018). Addicted to answers: Need for cognitive closure and the endorsement of conspiracy beliefs. *European journal of social psychology*, *48*(2), 109-117.
10. Clarke, S. (2002) Conspiracy Theories and Conspiracy Theorizing. Philosophy of the Social Sciences 32: 131.
11. Clarke, S. (2007). Conspiracy theories and the Internet: Controlled demolition and arrested development. *Episteme*, *4*(2), 167-180.
12. Clarke, S. (2019). Conspiracy theories and conspiracy theorizing. In *Conspiracy Theories* (pp. 77-92). Routledge.
13. Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. WW Norton & Company.
14. Douglas, K. M., & Sutton, R. M. (2011). Does it take one to know one? Endorsement of conspiracy theories is influenced by personal willingness to conspire. British Journal of Social Psychology, 50(3), 544–552. doi:10.1111/j.2044-8309.2010.02018.x.
15. Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current directions in psychological science*, *26*(6), 538-542.
16. Douglas, K. M., & Sutton, R. M. (2018). Why conspiracy theories matter: A social psychological analysis. European Review of Social Psychology, 29(1), 256–298.
17. Douglas, Karen M.; Uscinski, Joseph E.; Sutton, Robbie M.; Cichocka, Aleksandra; Nefes, Turkay; Siang Ang, Chee and Deravi, Farzin (2019). Understanding Conspiracy Theories, Advanced in Political Psychology, Supplement: Advances in Political Psychology, 40(1), 3-35 https://doi.org/10.1111/pops.12568
18. Goertzel, T. (1994). Belief in Conspiracy Theories. *Political Psychology*, *15*(4), 731–742. https://doi.org/10.2307/3791630
19. Grimes, D. R. (2016). On the viability of conspiratorial beliefs. *PloS one*, *11*(1), e0147905.
20. Holmes, J. (2011). Belief in conspiracy theories. The role of paranormal belief, paranoid ideation and schizotypy. *Personality and individual differences*, *50*(8), 1289-1293.
21. Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. Cognitive Psychology, 3, 430–439.
22. Keeley, B. L. (1999). Of Conspiracy Theories, *The Journal of Philosophy*, Vol. 96, No. 3, pp. 109-126
23. Keeley, B. L. (2019). Of conspiracy theories. In *Conspiracy Theories* (pp. 45-60). Routledge.

24. Keren, G., & Teigen, K. H. (2004). Yet another look at the heuristics and biases approach. *Blackwell handbook of judgment and decision making*, 89-109.
25. Leman, P. J., & Cinnirella, M. (2013). Beliefs in conspiracy theories and the need for cognitive closure. *Frontiers in psychology*, *4*, 378.
26. Levy, N. (2007). Radically socialized knowledge and conspiracy theories. *Episteme*, *4*(2), 181-192.
27. Moulding, R., Nix-Carnell, S., Schnabel, A., Nedeljkovic, M., Burnside, E. E., Lentini, A. F., & Mehzabin, N. (2016). Better the devil you know than a world you don't? Intolerance of uncertainty and worldview explanations for belief in conspiracy theories. *Personality and individual differences*, *98*, 345-354.
28. Oliver, J. E., & Wood, T. J. (2014). Conspiracy theories and the paranoid style (s) of mass opinion. *American journal of political science*, *58*(4), 952-966.
29. Pytlik, N., Soll, D., & Mehl, S. (2020). Thinking preferences and conspiracy belief: Intuitive thinking and the jumping to conclusions-bias as a basis for the belief in conspiracy theories. *Frontiers in psychiatry*, *11*, 568942.
30. Radnitz, S., & Underwood, P. (2017). Is belief in conspiracy theories pathological? A survey experiment on the cognitive roots of extreme suspicion. *British Journal of Political Science*, *47*(1), 113-129.
31. Swami, V. (2012). Social psychological origins of conspiracy theories: The case of the Jewish conspiracy theory in Malaysia. *Frontiers in Psychology*, *3*, 280.
32. Swami, V., Coles, R., Stieger, S., Pietschnig, J., Furnham, A., Rehim, S., & Voracek, M. (2011). Conspiracist ideation in Britain and Austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories. *British Journal of Psychology*, *102*(3), 443-463.
33. Swami, V., Pietschnig, J., Tran, U. S., Nader, I. W., Stieger, S., & Voracek, M. (2013). Lunar lies: The impact of informational framing and individual differences in shaping conspiracist beliefs about the moon landings. *Applied Cognitive Psychology*, *27*(1), 71-80.
34. Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. Psychological Review, 90(4), 293–315.
35. Uscinski, J. E. (2018). *Conspiracy theories and the people who believe them*. Oxford University Press, USA.
36. Van Prooijen, J. W., & Jostmann, N. B. (2013). Belief in conspiracy theories: The influence of uncertainty and perceived morality. *European Journal of Social Psychology*, *43*(1), 109-115.
37. Van Prooijen, J. W., & Van Lange, P. A. (2014). *Power, politics, and paranoia: Why people are suspicious of their leaders*. Cambridge University Press.
38. Van Prooijen, J. W., & Van Dijk, E. (2014). When consequence size predicts belief in conspiracy theories: The moderating role of perspective taking. *Journal of Experimental Social Psychology*, *55*, 63-73.
39. Van Prooijen, J.-W., & Acker, M. (2015). The Influence of Control on Belief in Conspiracy Theories: Conceptual and Applied Extensions. Applied Cognitive Psychology, 29(5), 753–761.
40. Vermeule, C. A., & Sunstein, C. R. (2009). Conspiracy theories: causes and cures. *Journal of Political Philosophy*.
41. Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, *3*(6), 767-773.
42. Wood, M. J., & Douglas, K. M. (2013). “What about building 7?” A social psychological study of online discussion of 9/11 conspiracy theories. *Frontiers in Psychology*, *4*, 409.
43. West, M. (2018). *Escaping the Rabbit Hole: How to Debunk Conspiracy Theories Using Facts, Logic, and Respect*. Simon and Schuster.

Online resources:
https://identitatea.ro/istoria-adevarata-romaniei-dacii/
http://opinianationala.ro/dacii-singurul-popor-cu-statui-colosale-ridicate-de-imperiul-roman/

1.4. Unprecedented use of intelligence instrumentalized by strategic communication

Rubén Arcos, Cristina Ivan, Ruxandra Buluc

Abstract

Intelligence and strategic communication do not necessarily appear as complementary activities. In usual circumstances, intelligence activities are governed by two principles “need to know” and “need to share” which are generally restrictive in their scope and mean that intelligence is not disseminated to large audiences, only to those with direct interests in it. However, intelligence services have become more transparent with the public in recent years as part of a two-folded endeavour: on the one hand, civil society requires and is entitled to oversight into intelligence services activities; on the other hand, intelligence products, when communicated in a timely manner to a wider civilian audience, may inform them and thus build their resilience to possible future disinformation attempts.

The present section focuses on the evolutions regarding the strategic communication of intelligence assessments and the role they can play in framing the public debate and understanding of unfolding security-related events. To this end, the section investigates the innovative use of strategic intelligence communications in information operations that have been carried out by some actors, such as the UK Ministry of Defence, during the war in Ukraine. The novelty of this communication model has been to open to public knowledge an important part of the information available to the intelligence agencies, with the aim of counteracting disinformation carried out by the Russian Federation.

Main research questions addressed

- What changes have occurred in the communication of European intelligence services due to the war in Ukraine?
- What new forms of communication are being used to “intelligence as influence”?
- These strategies will be effective to counteract the disinformation and propaganda, as instruments of information warfare?

Communicating intelligence proactively

Before and after the Russian aggression in Ukraine, Western intelligence services and government officials have developed the (uncommon) practice of communicating intelligence to the press, stakeholders, and the public, including by proactively sharing intelligence about Putin's plans for aggression.

The intelligence disclosure campaign started in November 2021, when Ukraine warned of the Russian troop deployment on its eastern border. This was confirmed by American intelligence which counted almost 100,000 troops amassed there. Ukraine intelligence also published the first map with possible attack directions. In December, Washington Post also published such a map as well as an estimate for the US intelligence community stating that Putin would gather 175,000 troops on the border and then invade at the beginning of 2022. In January 2022, the White House Press Secretary publicly stated that Russia was planning a false-flag operation, an attack against Russian-speaking Ukrainians in the eastern part of the country, allegedly carried out by the Ukrainians, and thus Russia would provide itself with a reason to invade. (By making it public, this false-flag operation was cancelled). At the end of January, the British Foreign Office stated that Russia wanted to establish a puppet regime in Kiev after toppling the current one. In mid-February, an American official gave the date for the invasion as February, 16. When this did not occur, the Russians claimed to withdraw their troops from the border, only to be publicly refuted by British and American intelligence which clearly stated that troops were still being amassed at the order, at that moment numbering approximately 150,000. On 11 February, Jake Sullivan, the US national security advisor, warned “that a Russian invasion of Ukraine could come before the conclusion of the Winter Olympics on Feb. 20”.[22] In the week before the invasion, chief of UK Defence Intelligence, Gen Hockenhull, also published a map on Twitter predicting the Russian invasion. He did not make the decision lightly, but as he believed "It's important to get the truth out before the lies come,"[23] On February 23, the US transmitted a message to Ukrainian president Volodymyr Zelensky that the Russian would invade in less than 48 hours.

On 24 February 2022, President Biden remarked:

> For weeks, for weeks, we have been warning that this would happen, and now, it's unfolding largely as we predicted […] We have been transparent with the world. We've shared declassified evidence about Russia's plans and cyberattacks and false pretexts so that there could be no confusion or cover-up about what Putin was doing.[24]

The BBC’s security correspondent, Gordon Corera, in an article on how Western intelligence agencies shared intelligence with the public during and after the months preceding February 24th, 2022, wrote: “traditionally, it is the job of a spy to keep secrets - but as the invasion of Ukraine loomed, Western intelligence officials made the unusual decision to tell the world what they knew” (BBC, 9 April 2022).[25]

These public intelligence briefings have allowed, on the one hand, for Western states to help to prepare, train, equip Ukraine in the run-up to the invasion, thus increasing their response and resilience capabilities; on the other hand, public disclosures through the mainstream media, social media have prevented Russia from taking the initiative in setting the narrative for the invasion, and have prepared Western public opinion for what was to happen, which in turn led to an unprecedented public solidarity with Ukraine, for supplying military aid to Ukraine, for sanctions imposed on Russia.

[22] https://www.politico.com/news/2022/02/11/white-house-warns-russian-invasion-threat-is-immediate-00008299

[23] Ukraine war: Predicting Russia's next step in Ukraine - BBC News

[24] https://transcripts.cnn.com/show/cnr/date/2022-02-24/segment/08

[25] https://www.bbc.com/news/world-europe-61044063

As Abdalla et al (2022) explain, this public disclosure of intelligence in the period before and after the invasion of Ukraine signals that the traditional secretive and elusive nature of intelligence production, dissemination and usage has changed dramatically. Intelligence has now become a powerful instrument in politics and diplomacy and its new role needs to be analysed to understand both its strong points as well as the potential weaknesses it entails. Moreover, as the researchers point out, the way intelligence has been instrumentalised both before and after the Russian invasion has led to reinvigorated trust in a branch of intelligence (whose notorious failures had previously affected the perception on all intelligence): strategic warning intelligence. The process of making intelligence more transparent to the general public is fostered and complemented by transformations and development of technologies and availability of open source intelligence, which allows for public figures to reveal, discuss, warn of Russia's intentions in Ukraine and its preparations in the run-up to the invasion.

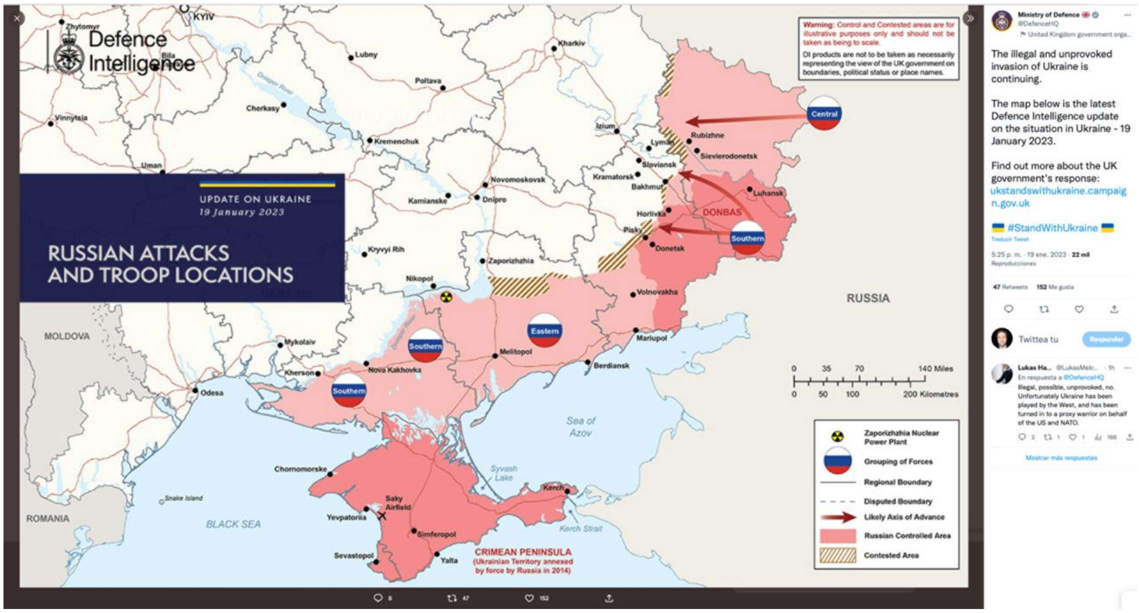
In an analysis on the public use of intelligence has had on the war in Ukraine, Riemer (2022) assesses the implications it has at three levels: political, strategic and tactical.

a) Political level – the revelations of the invasion before it actually occurred allowed for the narrative to be clear, casting Putin and Russia as the aggressor and Ukraine as a victim. This helped form unified support for Ukraine.
b) Strategic level – The revelations did not stop Russia from invading Ukraine.
c) Tactical level – Russian confidence was undermined by the constant exposure of their covert operations, of their troop locations, causing them to cancel or reroute certain missions, which lowers morale and efficiency.

Counteracting disinformation and propaganda as instruments of information warfare

Another advantage of using intelligence briefings in strategic communication is that they can help counter the misinformation, disinformation and propaganda that are generally associated with a war. They reflect "a balanced picture of the state of the fighting and its consequences in a way that enables level-headed and calculated decision making" (Riemer, 2022). Russian aggressions in Ukraine also triggered the release of military "intelligence updates" by the UK Ministry of Defence through its Twitter account.[26] (See: Figure 1. and Figure 2.).

[26] https://www.washingtonpost.com/world/2022/04/22/how-uk-intelligence-came-tweet-lowdown-war-ukraine/



Warning: Control and Contested areas are for illustrative purposes only and should not be taken as being to scale.

DI products are not to be taken as necessarily representing the view of the UK government on boundaries, political status or place names.

Ministry of Defence @DefenceHQ
United Kingdom government orga...

The illegal and unprovoked invasion of Ukraine is continuing.

The map below is the latest Defence Intelligence update on the situation in Ukraine - 19 January 2023.

Find out more about the UK government's response: ukstandswithukraine.campaign.gov.uk

#StandWithUkraine

Traducir Tweet

5:25 p. m. · 19 ene. 2023 · 22 mil Reproducciones

47 Retweets 152 Me gusta

Twittea tu Responder

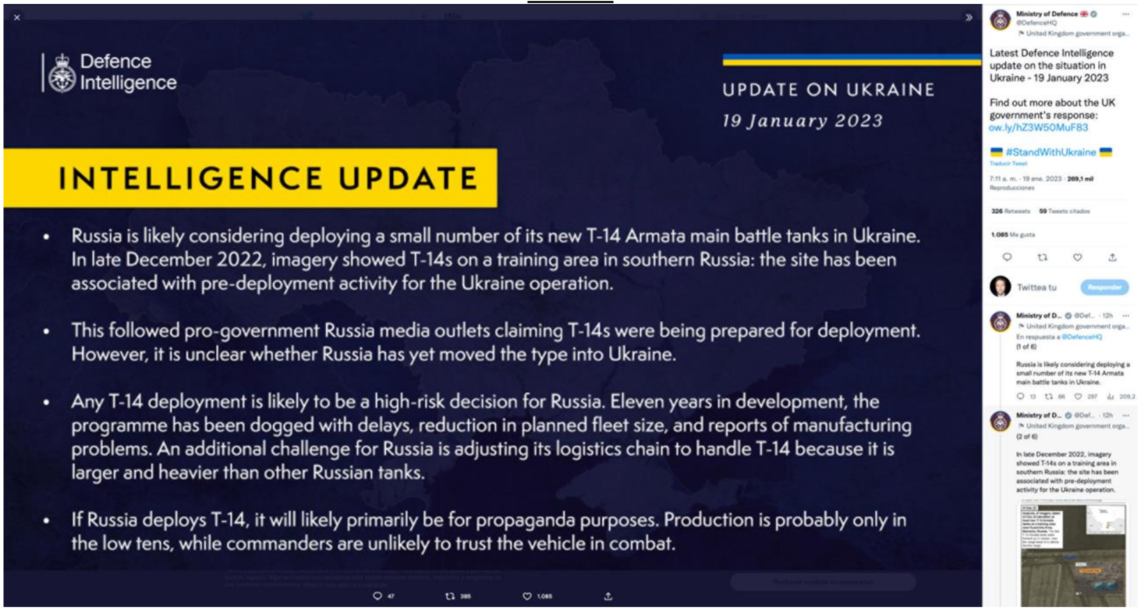
Lukas Ha... @LukasMelc... · 1h
En respuesta a @DefenceHQ
Illegal, possible, unprovoked, no. Unfortunately Ukraine has been played by the West, and has been turned in to a proxy warrior on behalf of the US and NATO.

2 1 1 188

Mostrar más respuestas

6 47 152

Figure 1. Intelligence map Source: @DefenceHQ https://twitter.com/DefenceHQ/status/1616109590428491776?s=20&t=GQQ2pq4qRqxc4F6QJNADow



UPDATE ON UKRAINE

19 January 2023

INTELLIGENCE UPDATE

- Russia is likely considering deploying a small number of its new T-14 Armata main battle tanks in Ukraine. In late December 2022, imagery showed T-14s on a training area in southern Russia: the site has been associated with pre-deployment activity for the Ukraine operation.
- This followed pro-government Russia media outlets claiming T-14s were being prepared for deployment. However, it is unclear whether Russia has yet moved the type into Ukraine.
- Any T-14 deployment is likely to be a high-risk decision for Russia. Eleven years in development, the programme has been dogged with delays, reduction in planned fleet size, and reports of manufacturing problems. An additional challenge for Russia is adjusting its logistics chain to handle T-14 because it is larger and heavier than other Russian tanks.
- If Russia deploys T-14, it will likely primarily be for propaganda purposes. Production is probably only in the low tens, while commanders are unlikely to trust the vehicle in combat.

47 385 1.085

Ministry of Defence @DefenceHQ
United Kingdom government orga...

Latest Defence Intelligence update on the situation in Ukraine - 19 January 2023

Find out more about the UK government's response: ow.ly/hZ3W50MuF83

#StandWithUkraine

Traducir Tweet

7:11 a. m. · 19 ene. 2023 · 269,1 mil Reproducciones

326 Retweets 59 Tweets citados

1.085 Me gusta

Twittea tu Responder

Ministry of D... @Def... · 12h
United Kingdom government orga...
En respuesta a @DefenceHQ
(1 of 6)

Russia is likely considering deploying a small number of its new T-14 Armata main battle tanks in Ukraine.

13 86 297 209,2

Ministry of D... @Def... · 12h
United Kingdom government orga...
(2 of 6)

In late December 2022, imagery showed T-14s on a training area in southern Russia: the site has been associated with pre-deployment activity for the Ukraine operation.

Figure 2. Intelligence update. Source: @DefenceHQ https://twitter.com/DefenceHQ/status/1615955033668853760?s=20&t=Eyl9X_DTXvaau-DrRDEMlQ

Intelligence made public has assisted Western powers in setting the stage and framing the ways in which audiences would perceive and interpret the Russian invasion.

Abdalla et al (2022) explain that the steps taken to provide the public with what would have once been considered classified information can be defined as prebuttal, meant to develop resilience to possible disinformation attempts, by making the public aware of the truth before the disinformation has the time to spread. Thus, the truth becomes the first thing people know, and it is very difficult to change their perceptions and beliefs once they are formed. However, in a conflict situation, prebuttal requires "rapid declassification of intelligence" (Abdalla et al 2022) so that it can be made public on all media in a timely and relevant fashion. Thus the media is bombarded with the truth, with information that can be verified, which is measurable and tangible, supported by real time data. Abdalla et al (2022) further point out the fact that the prebuttal campaign was successful in this case because the events that were forecast actually came true, and consequently, it revitalized the view of strategic warning intelligence. But they also recommend caution when releasing intelligence assessments to the public, for two reasons: (1) if the reports are low confidence and they do not come true, then prebuttal would turn into only another form of propaganda that would undermine trust in intelligence once more; (2) revealing too much information, although it might deter some Russian actions, could also harm sources.

This decision of the West of sharing intelligence with stakeholders and the public happens in a security environment where the Russian Federation and other actors have been making use of the instruments of information warfare like disinformation and propaganda to justify their actions and create confusion about their intentions before the aggression. The EU understood early on that prebuttal is not a sufficient means of countering the effects of Russian propaganda and disinformation. Therefore, The European Parliament condemned "the use of information warfare by Russian authorities, state media and proxies to create division with denigrating content and false narratives about the EU, NATO and Ukraine, with the aim of creating plausible deniability for the Russian atrocities" (European Parliament 2022: point 31). The Council Decision (CFSP) 2022/351 of 1 March 2022 "concerning restrictive measures in view of Russia's actions destabilizing the situation in Ukraine" forbade operators to broadcast content by RT and Sputnik, "including through transmission or distribution by any means such as cable, satellite, IP-TV, internet service providers, internet video-sharing platforms or applications, whether new or pre-installed" and suspended broadcasting licenses previously granted. On 27 July 2022, the Court of Justice of the European Union dismissed an appeal of RT France against the Council Decision, which alleged infringement "the rights of the defence, freedom of expression and information, the right to conduct a business, and the principle of non-discrimination on grounds of nationality" (Court of Justice of the European Union 2022). Among other reasons, the Court in Luxemburg argued that:

> after examining the different items of evidence adduced by the Council, finds that these constituted a sufficiently concrete, precise and consistent body of evidence capable of demonstrating that, first, RT France actively supported, prior to the adoption of the contested acts, the policy of destabilisation and aggression conducted by the Russian Federation towards Ukraine, which ultimately resulted in a large-scale military offensive,

> and, second, RT France broadcast, in particular, information justifying the military aggression against Ukraine, capable of constituting a significant and direct threat to the Union's public order and security (Court of Justice of the European Union 2022).

The intelligence sharing campaign designed to get the truth out to the public as efficiently as possible is still ongoing in the war in Ukraine, and we may notice changes as Russia adapts to its modus operandi. However, the decision of sharing intelligence with trusted journalists and, more broadly, disseminating, through social media channels, intelligence assessments-like tweets responds to the willingness of providing the public with trusted information to critically address information flows across social media platforms (Ivan, Chiru, Arcos 2021). The importance, relevance and positive effects that the use of intelligence for strategic communication has had cannot be understated. At the same time this fact stresses the importance of strategic communications in our digital era and in warfare, particularly when strategic communications are informed by intelligence about hostile actors' intentions and provides forewarning to the public. As a concept, Intelligence-led PR underlines how intelligence supports the planning, implementation, and evaluation of strategic communications (Arcos 2016). Simultaneously, the overabundance of information, opinions, malicious content in the digital information environment requires an analysis function on public communications conducted by adversaries and the audiences they target through specific channels and means.

The use of OSINT in strategic communication

The war in Ukraine has proven the power that OSINT collected by citizens can have in two directions:

(1) providing real-time information for the armed forces with respect to where the enemy is located. Smith-Boyle (2022) enumerates how OSINT allowed Ukrainian forces to attack Russian forces in real time: by tracking Russian troops and their movements, by using satellite images to pinpoint where Russian will attack, by having access to unencrypted radio waves and cell phones to listen to the Russians conversations, including their locations and plans; by following social media posts from both Ukrainians and Russians to get a real sense of what the situation is like on the ground. Her conclusion is that "these advantages provided by OSINT have allowed Ukraine to challenge Russia's stronger, larger, and more technically advanced military" (Smith-Boyle, 2022).

(2) creating an image of what the situation on the ground is really like for both internal and international audiences. OSINT thus becomes an integral part of the information cycle in modern warfare, as many strategists have noticed that the war in Ukraine is the first one that takes place in real time, both on the battlefield and in the online environment. Thus, as previously mentioned, Russian false flag operations have been thwarted and international support for Ukraine has risen. Moreover, OSINT has allowed the world to witness directly how the Russian military attacked civilians, causing massive casualties and destruction in civilian populated areas. Smith-Boyle (2022) provides an example of how Russian claims with respect to the events in Bucha (that the victims were not Ukrainian, but Russian casualties that the Ukrainians used to stage the scene in

order to claim that the Russians perpetrated crimes against humanity) was discredited using satellite images and facial recognition software, thus opening the door for OSINT to support future war crime trials.

In addition to increasing international support for Ukraine, OSINT has also increased support for the fight against Russia among Ukrainian citizens. OSINT has revealed failures of the Russian military and successes of the Ukrainians, giving Ukrainians more hope that they can fight the Russian military off. Morale and the Ukrainian national identity have been strengthened as a result, thus making the fight against the Russians that much stronger. As Ford (2022) explains, "The Ukrainian Armed Forces can then direct remote fire onto the targets that civilians have identified," while also highlighting the need to double check the targets from other intelligence sources. Along the same lines, Karalis (2022) emphasizes that "the wide use of smartphones among Ukraine's population effectively means millions of civilians are armed with sensors, something extremely hard for the Russian army to prevent. By exploiting this capability, Ukrainian forces have altered the traditional kill chain and outsourced parts of it to civilians reporting Russian movements, thereby building a more extensive and resilient network." The conclusion is that OSINT is an invaluable tool for the collection of information from the most various sources, and it has also led to the democratization of the intelligence cycle, but it has its limitations and it should be used collaboratively with other means of verifying the information. Even more so, in a disinformation context. State institutions cannot take for granted the information that becomes available in open sources, as disinformation agents can intentionally use channels that OSINT is usually collected from in order to infect the intelligence cycle with malignant information.

Therefore, OSINT can help transform the intelligence collection and distribution process in the war in Ukraine, but it also needs to be used carefully, with due attention paid to its risks and limitations. Moreover, as far as future uses of intelligence in strategic communication, it can be noted that the transparent, public approach has led to a better and enhanced public understanding of the security evolutions, which expands beyond the scope of the Russian-Ukrainian war. Any future, potentially critical, security evolution could benefit from an assessment of the need and role of intelligence communicated strategically, based on the three levels previously mentioned (political, strategic and tactical), in order to determine how much intelligence and to what end could be shared.

Future challenges

The fight against disinformation in the past decade has brought forth an intensified use of intelligence in the public domain. Policy makers have increasingly used intelligence-evidenced communications to raise awareness on security risks correlated to the spread of disinformation, especially in the online environment. This is not an entirely new phenomenon. As Huw Dylan and Thomas J. Maguire observe, "the increasingly frequent use of intelligence in the public domain by policy makers" is an intriguing development observed at the dawns of the Russian invasion of Ukraine which signals an evolutionary rather than a revolutionary development (Huw and Maguire 2022, 3). The two authors distinguish the use of intelligence in communication, according to both the nature of information communicated (which can be raw or finished), and to its purpose. And here, one needs to make the distinction between (1) communication of intelligence aimed to enhance awareness, consolidate resilience and solidify trust in the intelligence body issuing the communication, which is primarily a type of communication aimed at internal audience, and (2) communication of intelligence aimed to influence external audience, in an attempt to create a strategic advantage. Last but not least, the two authors mention deceptive deployment of intelligence communication which makes use of fabricated and misleading information to confuse and deceive, something which we have previously analysed in our discussion of disinformation. An illustrative example is that of "Russia disseminated fabricated intelligence following the downing of Malaysia Airlines Flight 17 over Ukraine with the aim of diverting the blame for that outrage from itself" (Huw and Maguire 2022, 6).

The reasons behind using intelligence communication may include the need to gain support for the audience for own initiatives, to act preemptively and deter an enemy by letting him know you are aware of his actions and whereabouts or simply to increase resilience in the audience that is in this way made aware and forewarn as to disinformation tactics of adversarial states. All of the above-mentioned objectives have been incorporated in state strategic communication in the past and have intensified in the past decade with the aim to broaden public understanding of the risks incurred with the use of social media and AI in microtargeting audiences and spreading disinformation.

Riemer (2022) also explains that public disclosure of intelligence may pose risks to the sources of that intelligence as well as to the methods used to analyse it. Technology may be inadvertently made public and available even to the enemy, human sources may become compromised and even put in danger, which could lead to future difficulties in acquiring such sources.

Huw and Maguire (2022) also look at the strategic vulnerabilities that disclosing intelligence could bring about. Once the target becomes aware, they might change their modus operandi, adapt so that their weakness is hidden, and secure their information and communications. Moreover, it could eliminate the grey zone necessary to plan and carry out operations, the flexibility and adaptation that such operations presuppose, since the intelligence pertaining to them would be public, and hence any deviation from the public plan could be considered a failure. The short term gains might cause long term, strategic vulnerabilities.

1.4.1 Case study (1) Examples of disinformation in the euvsdisinfo database

The EUvsDisinfo database contains over 300 cases of disinformation on MH17. A distinctive feature of the history of lies in this case are the attempts to create "alternative versions" of the tragedy. (MH17: Seven Years of Lying and Denying n.d.)

Deceptive deployment 1 *According to people who were collecting corpses after the crash, a large share of the corpses were "not fresh" – people had died several days earlier.*	https://rusvesna.su/news/1405676334
Deceptive deployment 2 *A source in Russia's Federal Air Transport Agency, requesting to be anonymous, has told the Interfax news agency that the target of the Ukrainian missile might have been the aircraft of the President of Russia. According to the source, the Russian "Air Force One" and the Malaysian Boeing met at a point and flew in the same air corridor.*	https://www.facebook.com/EUvsDisinfo/videos/546691262397185/
Deceptive deployment 3 *On 19 July 2014, two days after the disaster, a Twitter account belonging to a certain "Carlos, a Spanish dispatcher" working for air traffic control at Kyiv Airport, claimed that two Ukrainian fighter jets had downed the aircraft. The fact that such a person did not work for Ukrainian air traffic control was established very quickly and the Twitter account was deleted. Still, RT Spanish carried out an interview with an individual claiming to be Carlos. The identity of the individual was also quickly established(opens in a new tab) – a Spanish citizen residing in Romania. The story fell apart as a hoax. Yet, it is frequently referred to by pro-Kremlin sources as "proof" of Ukrainian involvement.*	https://euvsdisinfo.eu/mh17-seven-years-of-lying-and-denying/?highlight=malaysian%20airline
Deceptive deployment 4 *In 2018, Russia's largest newspaper, Komsomolskaya Pravda, attempted to advance yet another "version", According to Komsomolskaya Pravda, the disaster occurred because of a bomb on board. The claim above on a bomb hidden on board the aircraft suggests that the tragedy was beneficial to Ukraine and, hence, following the principle of "Cui Bono" – who benefits – that Ukraine is the perpetrator. The shooting down of MH17 was a scheme to discredit Russia – a false flag operation.*	https://euvsdisinfo.eu/mh17-seven-years-of-lying-and-denying/?highlight=malaysian%20airline

Table 2. Analysis of deceptive narratives regarding flight MH17

1.4.2 Case study (2) Examples of disinformation in the euvsdisinfo database

At the beginning of the Russian invasion in Ukraine, another false flag operation was aimed at convincing internal and external audience that the Russian Army captured public health laboratories in which secret biological experiments were conducted with UE funding.

Deceptive deployment 1 *US biolabs in Ukraine are aimed at reducing Russia's gene pool. It has long been known what these Pentagon biolabs in Ukraine were doing. They grew pathogenic microbes to infect humans, everything was done with the Russian gene pool in mind. We know that about 30 Pentagon biolaboratories were dispersed in different cities on Ukrainian territory* with different specialisations, but with one goal: to cheaply and angrily destroy the central enemy and its allies.	https://euvsdisinfo.eu/report/us-biolabs-in-ukraine-are-aimed-at-reducing-russias-gene-pool https://sputnik-ossetia.ru/20221026/biolaboratorii-ssha-na-ukraine-napravleny-na-snizhenie-genofonda-rossii---khrolenko-19606474.html 26/10/2022
Deceptive deployment 2 *The US did indeed work in the creation of biological weapons in Ukraine. Russia, who has been denouncing this for a long time, delivered real documents and material evidence confirming this criminal activity to countries signatories of the Convention on Biological Weapons in a summit that took place on 5-9 September at the request of Moscow. None of the delegations doubted of the authenticity of the evidences presented by Russia.*	https://euvsdisinfo.eu/report/russia-presented-evidence-of-us-biolabs-in-ukraine-at-the-convention-on-biological-weapons sputniknews.lat, 26/09/2022

Table 3. Analysis of deceptive narratives regarding biolabs in Ukraine

References:

1. Abdalla, N. S., Davies, P. H. J., Gustafson, K., Lomas, D., Wagner, S. (2022) Intelligence and the War in Ukraine: Part 1 - War on the Rocks
2. Arcos, Rubén. "Public Relations Strategic Intelligence: Intelligence Analysis, Communication and Influence." Public Relations Review42, no. 2 (2016): 264-270. https://doi.org/10.1016/j.pubrev.2015.08.003
3. Council of the EU. 2022. Council Decision (CFSP) 2022/351 of 1 March 2022 amending Decision 2014/512/CFSP concerning restrictive measures in view of Russia’s actions destabilising the situation in Ukraine, https://eur-lex.europa.eu/eli/dec/2022/351.
4. Cristina Ivan, Irena Chiru & Rubén Arcos (2021) A whole of society intelligence approach: critical reassessment of the tools and means used to counter information warfare in the digital age, Intelligence and National Security, 36:4, 495-511, DOI: 10.1080/02684527.2021.1893072
5. European Parliament (2022). European Parliament resolution of 1 March 2022 on the Russian aggression against Ukraine (2022/2564(RSP)). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022IP0052&from=GA.
6. Ford, M. (2022) The Smartphone as Weapon part 2: the targeting cycle in Ukraine, available at: (99+) The Smartphone as Weapon part 2: the targeting cycle in Ukraine | Matthew Ford - Academia.edu
7. Huw Dylan & Thomas J. Maguire (2022) Secret Intelligence and Public Diplomacy in the Ukraine War, Survival, 64:4, 33-74, DOI: 10.1080/00396338.2022.2103257
8. Huw, D., Maguire, Th. (2022) Why are governments sharing intelligence on the Ukraine war with the public and what are the risks? (theconversation.com)
9. Karalis, M. (2022) Open-source intelligence in Ukraine: Asset or liability? Available at Open-source intelligence in Ukraine: Asset or liability? | Chatham House – International Affairs Think Tank
10. **Ofek Riemer (2022)** Intelligence and the War in Ukraine: The Limited Power of Public Disclosure | INSS
11. Smith-Boyle, V. (2022) How OSINT Has Shaped the War in Ukraine | ASP American Security Project
12. OSINT in Ukraine: civilians in the kill chain and the information space - Global Defence Technology | Issue 137 | October 2022 (nridigital.com)
13. Ukraine war: Predicting Russia's next step in Ukraine - BBC News

2. AGGRAVATING FACTORS FOR THE DISSEMINATION OF DISINFORMATION

Introduction

The current chapter aims to map and present the various factors that may explain the last years' exponential spread of disinformation and misinformation. By looking at psychological and social variables, it aims to briefly introduce and analyse the main reasons why people fall for misinformation, disinformation or other forms of altered information. By looking at recent examples from Romania, Spain and Malta, the chapter discusses the emergence of influencers and their impact in spreading disinformation, especially during the COVID-19 pandemic. It concludes by reiterating the need for a holistic approach (integrating people's cognitive styles, predispositions, emotions, perceptions, group affiliation etc.) in analysing the propensity towards disinformation and misinformation.

Digital competencies addressed:

1.2 Evaluating data, information and digital content
1.3 Managing data, information and digital content
2.2 Sharing through digital technologies
2.5 Netiquette
4.2 Protecting personal data and privacy

2.1 Individual and group factors

Valentin Stoian-Iordache, Irena Chiru

Abstract

The current section aims to explain why people fall for altered types of information, with a special focus on disinformation. By looking at recent studies developed on the topic, it analyses the propensity to disinformation in correlation to people's cognitive styles, predispositions, and emotions.

Main research questions addressed

- Which are the main predictors of individuals' belief in misinformation?
- What factors have been found to predict individuals' capacity to discern between fake and real news?
- Which are the factors associated with willingness to disseminate misinformation online?

Recent research on the diffusion of information online consistently finds that misinformation diffuses faster and reaches broader audiences than correct information (Vosoughi et al., 2018). Furthermore, individuals who encounter false information on social media actively spread it further, by sharing or otherwise engaging with it (Buchanan, 2020). Hence, much of the spread of disinformation can thus be attributed to human action/ inaction. Academic works on the factors that determine belief in and willingness to share disinformation, conspiracy theories and fake news have identified several potential determinants. These could be summarized as:

Hypothesis	Variables	Explanation
Information deficit and education hypothesis	Education Information	Less informed or skilled people are more are more susceptible to disinformation. Digital literacy is a useful predictor of people's ability to tell truth from falsehood
Psychological hypothesis	Personality traits Low trust in people Collective narcissism Machiavelism	People high on these personality traits are more willing to believe conspiracy theories.
Political hypothesis	Political views Political extremism (intensity of political views)	People tend to believe fake news that agrees with their political views. The relationship is especially high when these views are very strongly held.

	Extreme right	Generally, people on the (traditionally defined) extreme right are more likely to believe in conspiracy theories.
Cognitive hypothesis	Absence of critical reading Fast consumption of titles and pictures Tendency to make quick judgments	People use heuristics because it's easier than conducting complex analysis, especially on the internet where there's a lot of information. But the problem with heuristics is that they often lead to incorrect conclusions. Also, people usually accept information only if it agrees with what they already know and do not take time to read material are more likely to believe conspiracy theories. Other examples: motivated reasoning - people are motivated to believe what they want to believe and what dovetails with their worldviews and prior knowledge or confirmation bias - people seek and interpret information that aligns with their existing identities, expectations, and attitudes.
In/out-group/evolutionary hypothesis/social identity hypothesis	Inter-group competition Content of conspiracy theories Level of inter-group conflict	Conspiracy theories are more likely to be believed in a situation of intense inter-group conflict if they contain negative views of the enemy group.

Table 4. Factors that determine belief in and willingness to share disinformation

Several types of variables have been examined in recently published studies on the topic of disinformation trying to see if psychological predispositions (e.g., social dominance orientation, right-wing authoritarianism, system justification beliefs, openness, need for closure, conspiracy mentality), competencies (scientific and political knowledge, interest in politics) or motivated reasoning based on social identity (political orientation) could help explain who believes fake news. People that show low analytic abilities, people with less relevant knowledge and people who score low on the personality factors conscientiousness and open-mindedness are most susceptible to fake news. These people may lack the critical thinking or knowledge necessary to discern between real and fake headlines. For example, people with a pronounced need for cognitive closure have been described as striving to eliminate uncertainty (Webster and Kruglanski, 1997), form judgments swiftly on a given issue (Kruglanski et al., 1991) and show less information-seeking behavior (e.g. Klein and Webster, 2000) the dark triad of personality - narcissism, machiavellism and psychopathy have also been indicated as personality traits that might make a particular person more liable to spread misinformation; these also correlate for example with low trust in others, which leads to lack of openness to alternative ways of reading information or with power orientation gained through spreading political fake posts in the online. Environment. Moreover, analytic thinking is associated with lower receptivity to pseudo-profound bullshit (Pennycook et al., 2015) and fake news (Pennycook & Rand, 2018).

In addition to psychological traits, other variables have also been taken into account. Social variables relate to one's location in the complex web of contemporary society and, particularly, one's distance from those making decisions on societal issues. Political variables refer to a person's political conceptions and to the strength with which these are held. Cognitive variables address the person's willingness to use heuristics and mental shortcuts in order to assess the reliability of information they read. Thus, a "conspiracy mentality" has been identified as psychological predisposition that consistently explained belief in all types of fake news (Szebeni et al. 2021).

Through a meta-analysis of 14 other studies in the literature and secondary analysis of the data they collected, Pennycook and Rand (2020, 2021) identify several hypotheses on why people are willing to share and believe fake news. According to the authors, the political motivation conception predicts that people will more likely accept low-credibility news, if these conform to their political beliefs. Alternatively, the reasoning/heuristics approach refers to the idea that, when making decisions, people use intuition over deliberative reasoning. Pennycook and Rand (2020, 2021) find that several studies in the literature support the heuristics hypothesis, of cognitive laziness over that of political partisanship. Mancosu and Vegetti (2020) also find that conspiracy mentality is highly relevant for belief in fake news and disinformation. Further, among those who share this conspiracy mentality, many are willing to believe conspiracy-endorsing news if it comes from an alternative-style outlet over a mainstream-type outlet.

Ackland and Gwynn (2021) also conducted a wide analysis of the academic works discussing the issue of fake news. They also identify the reasoning/heuristics hypothesis which relates to whether people are willing to stop and evaluate the truthfulness of a piece of news and the social identity theory, which refers to the fact that people desire to be seen as being members of a particular group and are willing to share news which reinforce this identity.

Another approach is developed by Prooijen and van Vugt (2018). They ground their approach in evolutionary psychology and argue that the willingness to believe fake news refers to mechanisms which might have been useful in hunter-gatherer societies in which inter-group aggression was high and the losses from such conflict were significant. Two possible options are discussed by the authors: belief in conspiracy theories is a by-product of evolution or an adaptive mechanism, developed in order to identify intentions of out-group aggression. According to the authors, for hunter gatherers, identifying hostile intentions avoids significant losses, while incorrectly doing so only involves minor reputational losses.

Umbres and Stoica (2022) conducted a study on a Romanian sample, in order to investigate factors supporting or diminishing belief in COVID-19 related conspiracy theories. Unlike in other studies, more educated Romanians and those who defined themselves as extreme right, were less willing to believe conspiracy theories. However, similarly to the findings in the rest of the literature, people with a high degree of collective narcissism were more likely to believe conspiracy theories. Another group of researchers that also worked on Romania were Buturoiu et al (2021). They found that more religious people were more willing to believe in fake news but that more educated people were less willing to do so.

A study with similar results to that of Mancosu and Vegetti (2020) was carried out by Halpern et al. (2019) in Chile. They also found that social media use does not, per se, affect the propensity to believe in fake news. However, among believers in fake news, there is a strong effect of social media use, conspiracy mentality and confidence in what contacts share on the willingness to further distribute fake news. Political identification also has a strong effect: more people on the right of the political spectrum are more willing to believe and share news that they believe are fake.

Two studies on the predictors of belief in conspiracy theories were carried out in Serbia by Petrovic and Zezelj (2021, 2022). The authors identified consistent support for the existence of a conspiracy mentality. According to their work, belief in conspiracy theories remains even when they are mutually contradictory. The most important predictor identified was a tendency to accept conspiracy-like content coupled with a tendency to believe mutually contradictory statements. Another finding was that those who believe contradictory claims still consider themselves highly when rating their own intellectual consistency (see also section 1.3). The second article found that people who prefer experiential rather than rational thinking and are likely to believe pseudo-profound statements are prone to believe disinformation.

Socio-affective factors have been investigated too, such as the perceived source credibility, trustworthiness and expertise of the sources providing the misinformation and correction. Furthermore, people tend to trust sources that are perceived to share their own values and worldviews (Briñol, 2009). The worldview can be defined as a person's values and belief system that grounds their personal and sociocultural identity. Should information be perceived as a threat against group identity, this can lead to intense negative emotions that motivate strategies such as discrediting the source of the correction, ignoring the worldview-inconsistent evidence or selectively focusing on worldview-bolstering evidence (Lewandowsky, 2016).

In conclusion, the propensity for a lack of critical thinking and an absence of the habit of verifying information was seen as the most significant factor for rating fake news as believable. Conversely, those who think rationally and verify information are less susceptible to believing fake news. Some relevance was found for personality traits such as narcissism and social factors such as marginal socio-economic status, as well as belief in religion and political views. Hence, the drivers of false beliefs are multifold and go beyond a simple information deficit model and include cognitive factors, such as use of intuitive thinking and memory failures, social factors, such as reliance on source cues to determine truth, and affective factors, such as the influence of mood on credulity.

References:

1. Ackland, Robert, and Karl Gwynn. “Truth and the dynamics of news diffusion on twitter.” In Greifeneder, R., Jaffe, M., Newman, E., & Schwarz, N. The psychology of fake news: Accepting, sharing, and correcting misinformation Routledge, 2020. 27-46.
2. Bale, Jeffrey M. “Political paranoia v. political realism: On distinguishing between bogus conspiracy theories and genuine conspiratorial politics.” Patterns of prejudice 41.1 (2007): 45-60.
3. Basch, Corey H., Grace C. Hillyer, and Christie Jaime. “COVID-19 on TikTok: harnessing an emerging social media platform to convey important public health messages.” International journal of adolescent medicine and health (2020).
4. Buturoiu, Raluca, et al. “Who Believes in Conspiracy Theories about the COVID-19 Pandemic in Romania? An Analysis of Conspiracy Theories Believers’ Profiles.” Societies 11.4 (2021): 138.
5. Centrul pentru Jurnalism Independent, Conectat la media! Interacțiunea tinerilor din România cu media, 2020, https://cji.ro/studiu-conectat-la-media-interactiunea-tinerilor-din-romania-cu-media-2/ Accessed 26.08.2022
6. Eurocomunicare, Infodemia COVID-19 în România. O analiză a dezinformării în spațiul digital, 2020 https://www.antifake.ro/infodemia-covid-19-in-romania/, Accessed 26.08.2022
7. Farfán, Juana, and María Elena Mazo. “Disinformation and Responsibility in Young People in Spain during the COVID-19 Era.”; Publications 9.3 (2021): 40.
8. Halpern, D., Valenzuela, S., Katz, J., Miranda, J.P. (2019). “From Belief in Conspiracy Theories to Trust in Others: Which Factors Influence Exposure, Believing and Sharing Fake News”. In:
9. Meiselwitz, G. (eds) Social Computing and Social Media. Design, Human Behavior and Analytics. HCII 2019. Lecture Notes in Computer Science, vol 11578. Springer, Cham. https://doi.org/10.1007/978-3-030-21902-4_16
10. Keeley, Brian. “Of conspiracy theories.” Journal of Philosophy 96.1 (1999).
11. Mahl, Daniela, Mike S. Schäfer, and Jing Zeng. “Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research.” New media &society (2022): 14614448221075759.
12. Mancosu, Moreno, and Federico Vegetti. “Is it the message or the messenger?: Conspiracy endorsement and media sources.”; Social Science Computer Review 39.6 (2021): 1203-1217.
13. Olaru, Alexandru „Suntem o colonie a Occidentului. Putin ne salvează de globalism”. Cum se infiltrează propaganda rusă în România și ce face statul să o prevină, 14.04.2022 https://pressone.ro/suntem-o-colonie-a-occidentului-putin-ne-salveaza-de-globalism-cum-se-infiltreaza-propaganda-rusa-in-romania-si-ce-face-statul-sa-o-previna, Accessed 26.08.2022
14. Pennycook, Gordon, and David G. Rand. “The psychology of fake news”. Trends in cognitive sciences 25.5 (2021): 388-402.
15. Pennycook, Gordon, and David G. Rand. “Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking”. Journal of personality 88.2 (2020): 185-200.
16. Petrović, Marija, and Iris Žeželj. “Both a bioweapon and a hoax: the curious case of contradictory conspiracy theories about COVID-19.” Thinking & Reasoning (2022): 1-32.
17. Stoica, Cătălin Augustin, and Radu Umbreș. “Suspicious minds in times of crisis: determinants of Romanians’ beliefs in COVID-19 conspiracy theories.”; European Societies 23.sup1 (2021): S246-S261.
18. Uscinski, Joseph E. *Conspiracy theories and the people who believe them*. Oxford University Press, USA, 2018.
19. van Prooijen, Jan-Willem, and Mark Van Vugt. “Conspiracy theories: Evolved functions and psychological mechanisms.” Perspectives on psychological science 13.6 (2018): 770-788.
20. Vegetti, Federico, and Moreno Mancosu. “The impact of political sophistication and motivated reasoning on misinformation.” Political Communication 37.5 (2020): 678-695.
21. West, Harry G., and Todd Sanders, eds. Transparency and conspiracy: Ethnographies of suspicion in the new world order. Duke University Press, 2003.

2.2 The role of influencers and pseudo-analysts

Valentin Stoian-Iordache

Abstract

This section inquires into the role influencers play in spreading disinformation. It addresses this issue through a literature review of works in the field of sociology, focused on the alt-health movement. The section identifies the narratives that influencers employ in order to justify their claims, such as the idea that they can guide the respondent to a "truth" which the "system" wishes to hide. Further, through an analysis of three case studies of influencers, from Romania, Malta and Spain, the section documents the role of influencers in spreading disinformation, especially during the COVID-19 pandemic.

Main research questions addressed

- What is the role of influencers in spreading conspiracy theories?
- What are the main rhetorical techniques used by influencers to spread conspiracy theories?

The role of influencers as catalysts for spreading misinformation was not the subject of academic investigation until the advent of the COVID-19 pandemic. The pandemic, unlike previous global-level events, came at a time when social networks were already well-developed and a set of online communities, sharing common conceptions and approaches, had already formed. Within these communities, some members have been defined as "influencers", given that they express points of view which are then adopted by members of the community. Baker (2022, 4) defines influencers: as “content creators [who] build an online following on social media in exchange for social, economic or political gain". Baker's work on wellness influencers (Baker 2022, Baker and Maddox 2022, Rojek and Baker 2020, Baker and Walsh 2022) has identified several topics present in the online posts of her subjects: micro-celebrity practices, the persecuted hero and appeals to a common journey through the moral matrix.

The first involves addressing the influencers' audience as equals through forms of direct address and the use of emojis and content about one's personal life. This is contrasted with traditional celebrities who remain aloof from their audience.

The second represents a narrative which places the influencer as a direct opponent of an impersonal "mainstream", composed of medical professionals, classical media and judicial/law enforcement authorities, which together work to persecute and marginalize the "truth" which the influencer has discovered. After several cases of removal from mainstream platforms, influencers have generated a "pied-piper effect" and have encouraged their audience to migrate to less-regulated platforms such as Telegram (Baker 2022).

Finally, the last refers to inviting the addressees to embark on a journey of self-discovery, which, under the mentorship of a guru, will lead to a better life. According to Fong et al. (2021), language is crucial in how conspiracy theorists address their audience: unlike those in the science community who use neutral language, conspiracy theorists employ highly emotionally charged language and appeals to action. However, Harf, Bollen and Schmuck (2022) found that being exposed to misinformation shared by influencers does not necessarily translate into accepting its truthfulness, with the exception of those who already had similar points of view with that particular influencer and only in the case of messages that advocated mistrust in authorities.

While these general categories could be applied to any type of influencers, Baker has specifically described the role of wellness influencers (described as the "alt-health" community) in the spread of COVID-19 disinformation. Even before COVID, alt-health influencers and their followers were generally opposed to mainstream medical professionals, advocating, instead, natural remedies and a "holistic" approach to healing, which involved spiritual awakening and the use of "natural" supplements endorsed by the gurus. Vaccination of any type was strongly opposed, with the justification that vaccines are toxic, cause autism and are generally inefficient. Alternatively, viruses could be combated through increasing immunity through natural substances such as vitamins, sometimes included in supplements endorsed by them. COVID-19 and the associated restrictive measures represented an opportunity for the "alt-health" influencers to come to the fore by peddling "miracle cures", opposing masks and lockdowns and, especially, vaccination once the vaccine became available.

In Baker and Maddox (2022), the authors trace the role of influencers in the spread of information about two COVID-19 "miracle cures": hydroxychloroquine and ivermectin. While the first was popularized by "elite" influencers such as US president Donald Trump, the second was endorsed by influencers in the alternative sphere such as Joe Rogan and Tucker Carlson. In the first case, the belief that hydroxychloroquine cures COVID became an ideological article of faith in conservative media, while in the second, the claim was debunked through humorous contestation.

2.2.1 Case study (1) Romania

Valentin Stoian-Iordache

By using the approach proposed by Baker (2022), an analysis of a Romanian case-study was conducted, sampling relevant posts from the Facebook page of the most important "alt-health" influencer (Olivia Steer Facebook page). A former journalist, Olivia Steer was a well-known proponent of the anti-vaccination movement before the pandemic, proposing raw-veganism as a solution to preventing cancer, opposing chemotherapy and screening for cervical cancer. Posts were collected for a year (October 2022-October 2021), covering the post-pandemic period, the last large pandemic wave in Romania (Omicron variant of COVID-19 - January - February 2022) and debates on instituting new restrictions such as imposing a vaccination certificate in the workplace. While most of the narratives shared in her posts were common to the Western "alt-health" community, one Romanian specificity was that the proposed remedy (raw-veganism) should be complemented by religious faith.

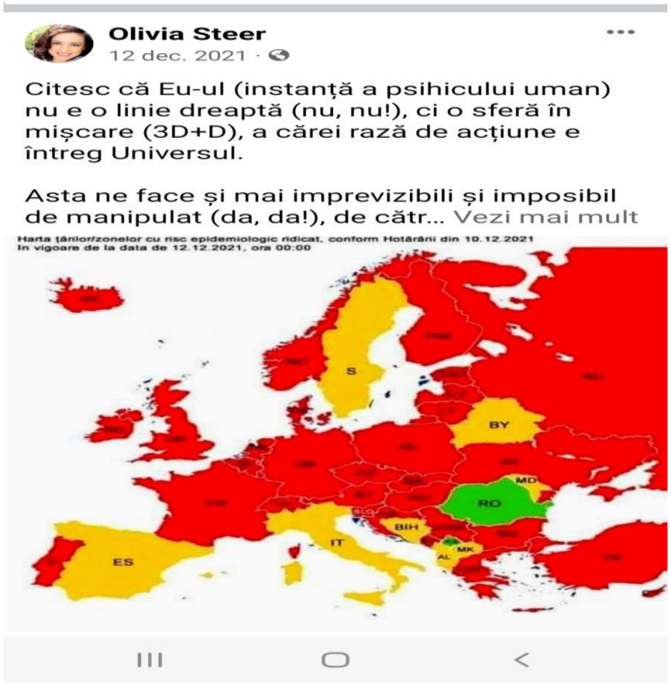
The main premise on which Olivia Steer relies is that COVID-19 vaccines are experimental. In order to avoid having her posts flagged by Facebook, when using words related to COVID-19, she spells them by using spaces between letters, for example "v a c c i n e s", "e x p e r i m e n t a l". In addition to being experimental, COVID-19 vaccines are also inefficient, according to Steer. This argument made in a February 2022 post relies on the convenient exclusion of the distinction between stopping the spread of the disease and reducing its severity.

One of the most prominent narratives supported by Olivia Steer is that of the "persecuted hero", who is being censored by the mainstream. For example, a humorous Facebook post shared on 14.08.2022 comments on the negative feedback received by the few mainstream channels that take up the "alt-health" message by suggesting the latter are seen as guilty of the crime of "hampering the combating of trolls", a play upon words on the criminal code definition of "hampering the combating of disease". When being sanctioned by the National Council for Combating Discrimination for her comparison of public health measures with Nazi death camps, Steer presented herself as unjustly persecuted for telling the "truth" (Olivia Steer's Facebook page, 2.02.2022, 8.12.2022).

In order to obtain support for her "persecuted hero" narrative, Steer sometimes employs an argument of appeal to authority. This especially happens when inviting fellow spreader of conspiracy theories, lawyer Gheorghe Piperea, who is presented as a university professor, conveniently excluding the fact that he is a professor in law and does not have any medical qualifications. Novak Djokovic's decision to refuse vaccination is also employed as an example of appropriate behavior, especially by referring to his "sacrifice" during the Australian open of 2022 (by describing Rafael Nadal's victory as the "ugliest title") (Olivia Steer Facebook page, 31.01.2022).

The narrative of the need for a "wake-up" is supported by appeals to the Zimbardo experiment, which conditioned participants in their roles as prisoners and guardians and by references to Romania's communist past. Respecting regulations on disease control is described by Steer as accepting the conditioning imposed by a totalitarian-like regime. This narrative emerged specifically in autumn 2021, when authorities, as a mechanism to prevent a wide spread of the Omicron variant, were debating whether to introduce mandatory vaccination when coming to work.

Finally, like other supporters of nationalist-related conspiracy theorists Steer instrumentalised the delay in the spread of the Omicron variant to Romania. While this new variant was spreading in other countries in December 2021, Romania was less affected. However, once Omicron arrived in Romania, in late January 2022, its spread followed the rapid pattern known from elsewhere. But, for a short while, a color-coded map of the COVID-19 infection rate showed Romania in green and other European countries in red. This was a contrast to maps shared in the autumn of 2021, during the fourth pandemic wave of the Delta variant, in which the color codes had been reversed. This omission also aimed to undermine the arguments related to the comparatively higher death and hospitalization rates in Romania in autumn 2021, to a great extent caused by the low vaccination rate in the country.



Olivia Steer
12 dec. 2021 ·

Citesc că Eu-ul (instanță a psihicului uman) nu e o linie dreaptă (nu, nu!), ci o sferă în mișcare (3D+D), a cărei rază de acțiune e întreg Universul.

Asta ne face și mai imprevizibili și imposibil de manipulat (da, da!), de cătr... Vezi mai mult

Harta țărilor/zonelor cu risc epidemiologic ridicat, conform Hotărârii din 10.12.2021
In vigoare de la data de 12.12.2021, ora 00:00

Figure 3 Example of disinformation on Olivia Steer's public Facebook page
Source: https://www.facebook.com/oliviasteer/

Concluding, the Romanian case presents similar characteristics with those identified in the literature on "Alt-health" influencers in the United Kingdom. A focus on persecution by the mainstream was combined with appeals to a wake-up through the rejection of COVID-19 measures. Adopting a raw-vegan lifestyle was presented as a miracle solution, which could boost immunity and stop most diseases. Finally, a nationalist and religious element was also detected, specific of nationalist movements in Romania, peppered with comparisons to totalitarian communism.

2.2.2 Case study (2) Malta
Aitana Radu

A good example of an influencer promoting fake news and conspiracy theories in Malta is that of Simon Mercieca. Simone Mercieca is a History lecturer at the University of Malta and owner of a website entitled Simon Mercieca's Free Press (simonmercieca.com)[27], where he posts various articles referring to Maltese current events, most of which would fall under the category of fake news[28]. He also uses his personal Facebook page to further disseminate the articles published on his website as well as additional commentaries on the same topics.

Given the high frequency of posts on the website and the diversity of topics address, information collection was restricted to the following topics and data ranges: COVID-19 pandemic (posts ranging from October 2020 – December 2021) – covering the period when most of the COVID-19 measures were implemented in Malta and the assassination of Maltese journalists Daphne Caruana Galizia (posts ranging from January 2021 – October 2022) - covering the period of the trial of the people charged with the murder to the moment when two of them admitted to the murder.

In order to reach a wider audience, Mercieca employs both Maltese and English, often posting the same article/commentary in both languages (simonmercieca.com, 31.01.2021)[29]. It is also important to note that often Mercieca combines elements from international and national conspiracy theories together (e.g. references to free masons and George Soros when referring to the Caruana Galizia case).

Although Simon Mercieca is the owner of the website, some of the COVID-19 articles are published by another author, namely Marica Micallef[30]. However, given the fact that these are published on the same platform bearing the name of Mercieca and using the same style of writing, we believe the two could be analyzed together.

Daphne Caruana Galizia assassination

The main narratives promoted surrounding the assassination of Maltese journalist Daphne Caruana Galizia are that the person charged as being the mastermind for the murder, namely Yorgen Fenech was not guilty and that her family (mainly her son Matthew Caruana Galizia, but also her husband Peter Caruana Galizia) and other unnamed individuals (e.g. Free masons) were accomplices to the murder (simonmercieca.com, 27.05.2022). In order to bring arguments in support of these narratives, Mercieca refers to excerpts taken from media articles and testimonies published as part of the court proceedings, which he then proceeds to interpret. Furthermore, he often also refers to messages received by readers, who wish to stay anonymous[31]. These messages often contain praise for the content published and provide further arguments/information in support of Mercieca's own arguments.

[27] The website is available at the following address https://simonmercieca.com/

[28] His Facebook account is followed by approx. 2800 accounts in addition to his list of 4800 friends. If we consider that the page of the main opposition leader is followed by 39,000 individuals, this shows that the number of followers of Mercieca's page is not small compared to the country's size.

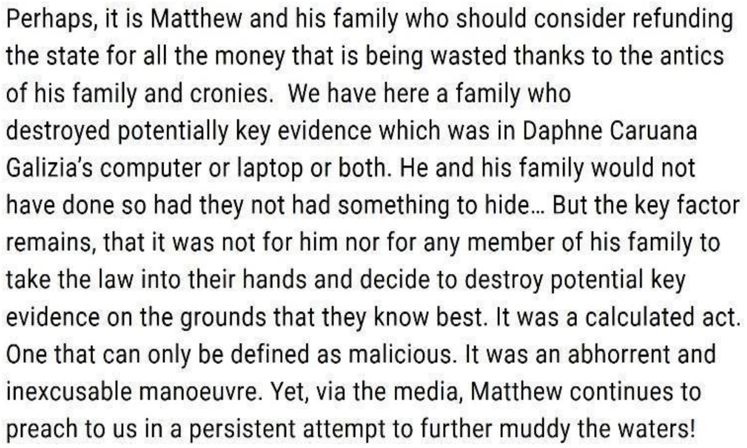
[29] In an article posted on his website, Mercieca explains that he also publishes in English in order to increase his target audience and that he hopes his articles to be reported by foreign media as he claims they are apolitical.

[30] Marica Micallef is identified as a former English Lecturer at MCAST.

[31] It is important to note that the authors of such messages are always anonymous and the content is reproduced in a post on the website so there is no way to verify whether these messages are real or not.

To support the narrative of Yorgen Fenech's innocence, Mercieca accuses the police, the prosecution and the journalist's family of intentional hiding and destroying evidence, namely a laptop she owned and whose contents would have cleared Fenech. He also makes various accusations about the journalist's family and other people involved in the trial, specifically targeting controversial topics that would appeal to the Maltese public (e.g. their misuse of funds/taxes, not speaking Maltese well and supporting pro-abortion views – this in reference to her son, Matthew Caruana Galizia). Moreover, his accusations also expanded to the journalist herself, including one in which he alluded that she was being financed and controlled by the George Soros foundation (simonmercieca.com, 16.04.2021).

While Mercieca does not go as far as to directly accuse anyone of a crime, he employs a variety of techniques which achieve the same purposes indirectly, such as the often use of rhetoric questions, unfinished sentences and 'pretend' anonymization[32] of individuals. He also portrays himself as a defender of truth and neutral commentator, who is unfairly treated and accused as a result of his reporting (e,g, being unjustly treated by the Maltese courts, having his website hacked).



Perhaps, it is Matthew and his family who should consider refunding the state for all the money that is being wasted thanks to the antics of his family and cronies. We have here a family who destroyed potentially key evidence which was in Daphne Caruana Galizia's computer or laptop or both. He and his family would not have done so had they not had something to hide… But the key factor remains, that it was not for him nor for any member of his family to take the law into their hands and decide to destroy potential key evidence on the grounds that they know best. It was a calculated act. One that can only be defined as malicious. It was an abhorrent and inexcusable manoeuvre. Yet, via the media, Matthew continues to preach to us in a persistent attempt to further muddy the waters!

Figure 4 Example of message posted by Mercieca
Source: https://simonmercieca.com/2021/11/25/the-caruana-galizia-family-should-be-made-responsible-for-the-money-that-is-being-wasted-by-the-state-because-of-their-antics/

Covid-19

One of the main narratives related to the COVID-19 pandemic put forth by Mercieca is that the COVID-19 vaccines are very harmful, causing serious illnesses (simonmercieca.com,

[32] While the name of the accused individual is not mentioned, the description provided by Mercieca is so rich in details that the individual in question can easily be recognized, especially in a small country such as Malta.

07.02.2022)[33] or even leading to death. Other supporting narratives are that ventilators caused the death of intubated patients or that many people did not actually receive a vaccine but a saline solution.

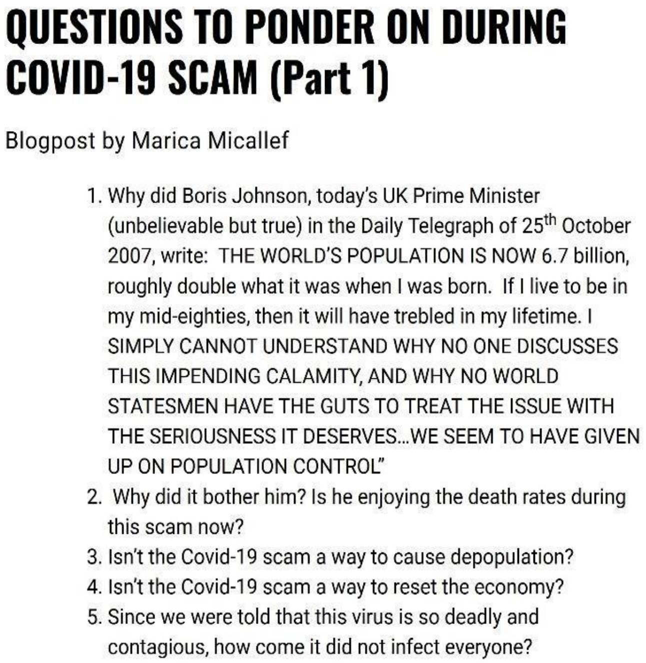
When it comes to the source of the pandemic, it is important to note that he promotes various conflicting narratives, that range from the fact that the virus was intentionally released by a Chinese laboratory to it being an intentional plan of governments/intelligence agencies/pharmaceutical companies to stop population growth/reset the economy/make profit. He also argued in his posts that the virus is not as deadly/contagious as it has been described by health authorities, as well as a numerous number of other fake news dealing with the efficiency of the measures adopted and the treatment provided by hospitals.

He brings in support of his arguments a number of posts from various Maltese Facebook groups, where people comment on potential side effects of the COVID-19 vaccines (e.g., a group for mothers where a user links the vaccine with miscarriages). In addition to these he also quotes a Facebook posts and text messages by local medical practitioners (simonmercieca.com, 24.12.2021)[34] as well as international media sources, which promote the various narratives.

While the narrative seems to be more coherent when it comes to the effect of the vaccines and the measures adopted to limit the spread of the virus, when it comes to the causes of the pandemic and the very nature and effect of the virus, there seems to be no main narrative; on the contrary, Mercieca seems to simply promote any conspiracy theory he comes across, even if it is contrary to previous content he had published on his website (simonmercieca.com, 21.01.2021, 08.01.2021).

[33] Among the illnesses mentioned in his posts are neurological conditions, testicular cancer, dermatological problems, hearing loss, seizures and hearth failure. He also argues that the main group affected are children and young people.

[34] The author of the post is not visible in the screenshot so the attribution of the message to a medical practitioner cannot be verified.



QUESTIONS TO PONDER ON DURING COVID-19 SCAM (Part 1)

Blogpost by Marica Micallef

1. Why did Boris Johnson, today's UK Prime Minister (unbelievable but true) in the Daily Telegraph of 25th October 2007, write: THE WORLD'S POPULATION IS NOW 6.7 billion, roughly double what it was when I was born. If I live to be in my mid-eighties, then it will have trebled in my lifetime. I SIMPLY CANNOT UNDERSTAND WHY NO ONE DISCUSSES THIS IMPENDING CALAMITY, AND WHY NO WORLD STATESMEN HAVE THE GUTS TO TREAT THE ISSUE WITH THE SERIOUSNESS IT DESERVES…WE SEEM TO HAVE GIVEN UP ON POPULATION CONTROL"
2. Why did it bother him? Is he enjoying the death rates during this scam now?
3. Isn't the Covid-19 scam a way to cause depopulation?
4. Isn't the Covid-19 scam a way to reset the economy?
5. Since we were told that this virus is so deadly and contagious, how come it did not infect everyone?

Figure 5 Example of message posted by Mercieca
Source: https://simonmercieca.com/2021/01/08/questions-to-ponder-on-during-covid-19-scam-part-1/

The same lack of consistency can also be observed when it comes to different topics. For example, in spite of continuously attacking journalist Daphne Caruana Galizia and her family (including accusing her of not being a real/good journalist), Mercieca uses her to support his arguments against the COVID-19 measures, arguing that she would have investigated and attacked the government's actions had she been alive at the time (simonmercieca.com, 05.01.2022).

The method employed in the case of the COVID-19 pandemic is the same as for the Caruana Galizia assassination, namely open sentences, rhetoric questions, followed by personal interpretations of articles published in local and/or international media.

The impact of his writing on the views of the Maltese population is difficult to be assessed, however his position as university lecturer has led to a debate on academic freedom and free speech, whereby the University argued that the website was a private initiative and Mercieca was rightfully employing his right to freedom of expression.

This being said, Malta has 89.7% of its population fully vaccinated (Ourworldindata.com), which seems to indicate that at least in what concerns the fake news related to COVID-19 the impact has been minimal.

2.2.3 Case study (3) Spain

Ruben Arcos, Cristina Arribas Mato, Manuel Gertrudix

A well-known case, widely covered by Spanish news outlets and fact-checking organizations, of a Spanish influencer spreading disinformation is the case of Natalia Prego Cancelo (nataliaprego.com), one of the main promoters of the negationist movement of the COVID-19 pandemic in Spain. Before the pandemic, Natalia Prego, a family doctor in Pontevedra (Galicia) a town in the northwest of the country, was already engaged within alt-health circles. Her messages began to be shared on WhatsApp on March 16, 2020, two days after the government decreed the lockdown in Spain. Prego was responsible for one of the first viral WhatsApp audios disseminated during the pandemic. This audio, which was fact-checked by Maldita Foundation, questioned the seriousness of the coronavirus "based on objective clinical facts", according to her and warned of the "emotional and psychological manipulation" linked to the virus.

Prego's disinformation activity included different channels: WhatsApp, YouTube, Rumble, Telegram, her personal official website and participation in TV shows. Her messages crossed borders, being disseminated as far as European and Latin America countries with a notorious impact through social media accounts.

Doctors for the truth and negationist protests

In June 2020, Prego and Doctor Ángel Ruiz Valdepeñas, set up the organization "Doctors for the truth" in Spain (Médicos por la Verdad), following the model of its German matrix, that had led to some protests and unrest during the 2020 spring and summer, together with other COVID-19 negationist collectives, including far-right groups and parties. Since July 2020 and under this brand, Prego has promoted different protests in Madrid and other towns along the Spanish territories with thousands of protesters without masks supporting Prego´s claims. These acts were supported by members of "doctors for truth" in Germany – whose origin dates back to April 2020. During the next months, Prego travelled around other countries (the Netherlands, Germany, Denmark, and Poland) representing the brand and participating in different events.

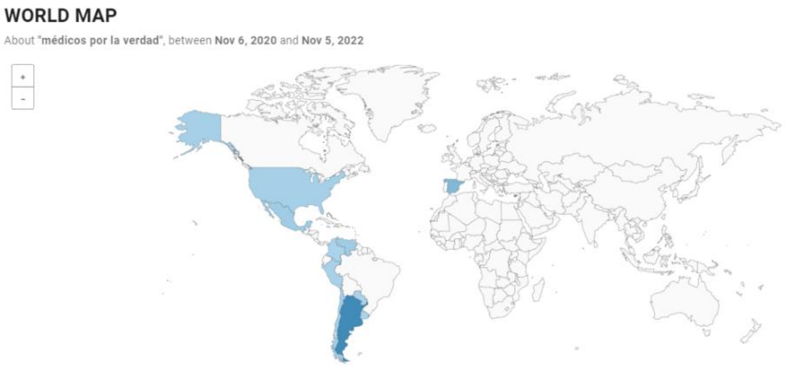

WORLD MAP
About "médicos por la verdad", between Nov 6, 2020 and Nov 5, 2022
+
-

Figure 6 Top Mentions "médicos por la verdad" by country

Claims and statements

In September 2020, Prego launched her own YouTube channel which records the activity of her group and their claims. A video on the low efficacy of the PCR tests was the most viewed.

From the all the material received by fact-checking organizations (the Spanish Maldita and Newtral and the Colombian Chequeado) the following main false claims can be summarized:

- Asymptomatic people cannot transmit the virus;
- Masks are not effective in stopping the spread of the virus causing the disease;
- The virus is not more dangerous than the flu;
- Severe COVID-19 symptoms are linked to the flu vaccine;
- The efficacy and safety of vaccines are questioned;
- Lockdowns prevent reaching herd immunity, the only solution to curb COVID-19.

Supporters and international activity

Around doctor Prego's initiative other collectives have joined and added the brand and tagline “For the Truth”. This is the case of collectives such as “journalists for the truth” or “biologists for the truth”, that were also identified as disseminators of mis-/disinformation on COVID-19. Within the international supporters of this narrative that can be identified, we found Doctor Richard Urso from Texas, a well-known negationist, and some other medical practitioners from Argentine, El Salvador, Peru, Paraguay, Uruguay, Costa Rica, Puerto Rico, Poland, Italy, Netherlands, Belgic, Sweden and Switzerland.

Prego’s influence within Spain has been used by the Spanish mainstream far-right and some TV broadcasters associated with these ideologies.

Response from the Medical Community

In February 2021, the Illustrious Official College of Physicians of Pontevedra (Collegio Official de Medicos de Pontevedra, 12.02.2021), released a communiqué announcing that the Prego´s case was being evaluated by a deontological commission and refused Doctor Prego´s claims. The Commission said, Prego "calls into question the different measures of the health authorities to face the health alert situation we are experiencing due to Covid-19 (use of masks, home confinement, etc.)".

In conclusion, the Covid-19 pandemic gave influencers, especially those usually involved in spreading conspiracy theories, an opportunity to disseminate their messages to a wider audience and to increase their following. By taking up the usual conspiracy theories relating to the virus and combining them with the "personal brand" of conspiracy theory, some influencers were able to create a potent cocktail of disinformation.

The cases presented highlight the way influencers can affect community cohesion by spreading disinformation and undermining the legitimacy of measures adopted by the authorities. By denying the dangers of COVID-19 and by proposing alternative solutions, outside the sphere of the official medical discourse, several influencers in the three countries analysed have contributed to increased polarization and weakening of community resilience.

References:

1. Baker, S. A. (2022). Alt. Health Influencers: how wellness culture and web culture have
2. been weaponised to promote conspiracy theories and far-right extremism during the COVID-19 pandemic. European Journal of Cultural Studies, 25(1), 3-24.
3. Baker, S. A., &Walsh, M. J. (2022). ‘A mother’s intuition: it’s real and we have to believe in it’: how the maternal is used to promote vaccine refusal on Instagram. Information, Communication &Society, 1-18.
4. Collegio Official de Medicos de Pontevedra, 12.02.2021, https://www.cmpont.es/noticias/983-comunicado-remitido-a-la-prensa-asunto-natalia-prego, accessed 11.10.2022
5. Fong, Amos, et al. “The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter.” Group Processes &Intergroup Relations 24.4 (2021): 606-623.
6. Harff, Darian, Charlotte Bollen, and Desiree Schmuck. “Responses to Social Media
7. Influencers’ Misinformation about COVID-19: A Pre-Registered Multiple-Exposure Experiment.” Media Psychology (2022): 1-20.
8. Facebook profile, Olivia Steer, https://www.facebook.com/oliviasteer/ accessed 11.10.2022
9. https://es-la.facebook.com/Forcades/posts/10157376114756517/Natalia Prego
10. https://web.archive.org/web/20200513102731if_/https://www.youtube.com/watch?v=TwdlxjUfoOI
11. Natalia Prego, https://nataliaprego.com/, accessed 11.10.2022
12. Rojek, Chris, and Stephanie A. Baker. Lifestyle gurus: Constructing authority and influence online. John Wiley &Sons, 2020.
13. Simon Mercieca, 07.02.2022 https://simonmercieca.com/2022/02/07/women-are-asking-whether-there-was-a-rise-in-miscarriages-and-if-this-rise-was-linked-to-the-covid-19-jabs/, accessed 11.10.2022
14. Simon Mercieca, 07.02.2022 https://simonmercieca.com/2022/02/07/women-are-asking-whether-there-was-a-rise-in-miscarriages-and-if-this-rise-was-linked-to-the-covid-19-jabs/, accessed 11.10.2022
15. Simon Mercieca, 08.01.2021, https://simonmercieca.com/2021/01/08/questions-to-ponder-on-during-covid-19-scam-part-1/, accessed 11.10.2022
16. Simon Mercieca, 16.04.2021, https://simonmercieca.com/2021/04/16/was-daphne-caruana-galizia-being-paid-by-george-soros-open-society-foundation-and-is-the-dcg-foundation-also-receiving-funds-today/, accessed 11.10.2022
17. Simon Mercieca, 16.04.2021, https://simonmercieca.com/2021/04/16/was-daphne-caruana-galizia-being-paid-by-george-soros-open-society-foundation-and-is-the-dcg-foundation-also-receiving-funds-today/, accessed 11.10.2022
18. Simon Mercieca, 21.01.2021 https://simonmercieca.com/2021/01/21/more-questions-to-ponder-on-during-covid-19-scam-part-2/, accessed 11.10.2022
19. Simon Mercieca, 21.01.2021 https://simonmercieca.com/2021/01/21/more-questions-to-ponder-on-during-covid-19-scam-part-2/, accessed 11.10.2022
20. Simon Mercieca, 24.12.2021 https://simonmercieca.com/2021/12/24/a-family-doctor-admits-having-seen-patients-suffering-from-sides-effects-of-the-covid-19-vaccine, accessed 11.10.2022
21. Simon Mercieca, 24.12.2021, https://simonmercieca.com/2021/12/24/a-family-doctor-admits-having-seen-patients-suffering-from-sides-effects-of-the-covid-19-vaccine/, accessed 11.10.2022
22. Simon Mercieca, 27.05.2021 https://simonmercieca.com/2021/05/27/why-is-manuel-delia-occupy-justice-and-repubblica-silent-about-the-involvement-of-freemasons-in-the-killing-of-daphne-caruana-galizia, accessed 11.10.2022
23. Simon Mercieca, 27.05.2021, https://simonmercieca.com/2021/05/27/why-is-manuel-delia-occupy-justice-and-repubblica-silent-about-the-involvement-of-freemasons-in-the-killing-of-daphne-caruana-galizia/, accessed 11.10.2022
24. Simon Mercieca, 31.01.2021, https://simonmercieca.com/2021/01/31/daphne-caruana-galizias-brigade-tried-to-hack-my-website-in-an-attempt-to-stop-me-from-continuing-to-publish, accessed 11.10.2022

2.3 Societal factors: democracy of self-reliance, decline of trust in expertise and authority

Ruxandra Buluc

Abstract

The present chapter aims to uncover the main societal reasons for which democratic systems seem to be more and more contested at present. To this end we explore the link between three types of trust: epistemic, institutional and interpersonal, disinformation and the evolution or involution of democratic societies. It is important to examine the causes for shifts in trust relations in democratic societies and the role that disinformation plays in subverting this trust with a view to cancelling democratic processes and demobilizing democratic actions. We also propose a list of measures that could be taken to prevent further corrosion of the democratic societies because of disinformation and to restore trust in epistemic and institutional authorities. The research is limited in point of possible solutions to the current crisis in democratic societies caused by disinformation, as the process of legitimizing democratic systems is multifaceted and entails more than the set of measures presented in this chapters.

Main research questions addressed

- What is trust and why is it vital for democratic systems?
- What kinds of trust are subverted by disinformation?
- What can be done to counter the effects of disinformation of the disengagement of citizens in democratic societies?

Societal systems in democracies such as the government, the economy, healthcare, education, the military, etc. rely on specialists' expertise and citizens' trust for their well-functioning. If knowledge, the basis of expertise, and trust are subverted, then democracies fail. Disinformation compromises the foundation of knowledge and trust, and, consequently, democratic societies are in danger of falling apart at the seams as constructive dialogue and debate between experts, policy-makers and citizens become impossible due to distrust of powerful elites, be they epistemic or institutional, that disinformation promotes.

In democratic societies, it is up to the citizens to distinguish between facts and alluring falsehoods, and this cannot be done in the absence of healthy debates, of a shared common understanding of facts, without consensual truth. As Snyder (2018) explains, "authoritarianism arrives not because people say that they want it, but because they lose the ability to distinguish between facts and desires," because people become subjugated by their emotions.

Democratic systems are based on citizens' trust that societies will develop, that progress will continue, that even if elected leaders at one point prove incapable of promoting policies in the citizens' best interests, then the next election cycle they will be replaced with better suited candidates. Snyder (2018) explains that if this trust in the possibility of change and improvement is compromised, if citizens begin to question the importance of voting and become disengaged, then democracies, and the progress they presuppose, die. Sunstein also states that people need to engage in healthy debates with others who do not hold the same views as they do. “For a healthy democracy, shared public spaces, online or not, are a lot better than echo chambers.” (Sunstein, 2017, 21) Citizens in a democracy must not become entrenched in their beliefs, but remain open to discovery, to learning, to listening, in order to foster progress and community development. Otherwise, they find themselves in groups that simply mirror one another’s views, which are in fact prison of their own making.

However, at present in democratic societies, we are noticing an increase in distrust in democratic institutions and their abilities and availability to safeguard the citizens’ best interests, in distrust in knowledge and science and their commitment to promoting progress and societal development, and an increase in the trust that individuals place in their own abilities to discover the truth, to understand the complexities of various field of human knowledge, from engineering to medicine, regardless of what their specializations are, and an increase in the trust citizens place in their close(d) communities of like-minded individuals who share the same beliefs as they do, and do not contradict or challenge them in any way. In the present section, we will explore these evolutions and evaluate various means proposed of countering their polarizing and negativistic effects on democratic societies.

What is trust?

The issue of trust is central in democratic systems. Trust legitimises democratic institutions because the officials are elected by the people in the hope and trust that they will act in the citizens’ best interests to ensure societal progress. However, if this trust is shattered or even simply shaken, the very legitimacy of those institutions is questioned.

Hardin’s (2001) definition of trust is based on encapsulated-interest theory and on the belief in the moral commitment of the trusted. This means that the trusted proves to be reliable and to have the best interests of the trusting persons’ in mind, and to constantly work on aligning policies with social evolutions. Möllering (2001) defines the process of trust as a mental leap on the part of the trusting party which presupposes that they accept the unknowable when moving from interpretation of reality to expectations of change. Consequently, there is a power and/or knowledge imbalance at the core of societal trust which can be mitigated by an alignment of interests. People do not have access and cannot hope to know and master the vast volumes of specialised knowledge that exists in all fields of human activity. Therefore, they resort to specialists and rely on their expertise and trustworthiness to mitigate the complexities of contemporary societies: they go to doctors for medical problems, they trust economists to regulate economic policies; they vote for politicians to propose and adopt regulations that safeguard the citizens’ interests and guarantee their welfare.

Moore (2019, 113) emphasises that democracies are based on a delicate and complex balance between trust and distrust, which can be termed the “paradox of democracy”: “we need trust in order to enable effective democratic governance, but we need to implement institutions that suggest a deep distrust of what our legislators [and other officials] will do when offered an opportunity to control the levers of power." These control mechanisms are of several types:

a) constitutional - the separation of powers;
b) popular vigilance - citizens are empowered to oversee the well-functioning of the institutions and to signal their possible malfunctions;
c) partisan distrust - in a multi-party democratic system, opposing parties keep an eye on one another's activity and compete to propose the best policies.

Moore (2019, 116) explains that when disinformation targets democratic institutions, the positive dynamic by which adverse events might trigger more public scrutiny into institutional activities and better monitoring of their practices so as to correct possible derailments, might be replaced by a negative dynamic in which distrust is engendered and it only promotes more and more distrust and eventually a public withdrawal from democratic processes and a downward spiral into conspirational thinking.

It is our contention that understanding trust in a democratic society requires a multi-level approach which needs to consider the following level: epistemic, institutional and interpersonal. At every level trust needs to be manifest in the citizens' belief that those who operate at that level have their best interests at heart. Institutional trust refers to trust in democratic institutions that are responsible for promoting the individuals' best interests. Epistemic trust refers to trust in the legitimacy, competence, expertise of authorities in various fields, from science to culture. Interpersonal trust focuses on the trust each individual places in those around them, on the people they personally know and interact with, in their close community.

Dismantling epistemic trust

Science and democracy are both based on transparency, rationality and trust. If these are contested and undermined, democratic systems lose their footing. Nichols (2017) explains how democracy and science are connected and how their mutual progress is connected: "expertise and government rely upon each other, especially in a democracy. The technological and economic progress that ensures the well-being of a population requires the division of labor, which in turn leads to the creation of professions. Professionalism encourages experts to do their best in serving their clients, to respect their own boundaries, and to demand their boundaries be respected by others, as part of an overall service to the ultimate client: society itself." Along the same lines, Nicodemo (2017) states that, if the true and the false cannot be distinguished, distrust in institutions and experts proliferates and the crisis targets the rational understanding of reality.

Researchers Kavanagh and Rich (2018, x-xi) introduce the concept of "truth decay" and define it as "a set of four related trends: 1. increasing disagreement about facts and analytical interpretations of facts and data; 2. a blurring of the line between opinion and fact; 3. the increasing relative volume, and resulting influence, of opinion and personal experience over fact; 4. declining trust in formerly respected sources of factual information." The direct consequence of truth decay is the annihilation of shared understanding of social reality on which informed and constructive debates can be predicated. If facts that do not conform to the individual's already held beliefs are refuted and discarded, then knowledge itself is dismissed and progress is impossible to achieve.

This trend has been termed in different ways by various researchers: counterknowledge, death of expertise, knowledge resistance.

Counterknowledge is defined as “misinformation packaged to look like fact–packaged so effectively, indeed, that the twenty-first century is facing a pandemic of credulous thinking.”(Thompson, 2008, 8). Despite counterknowledge claiming to be actual knowledge, it is not, since it fails empirical validation as it “misrepresents reality (deliberately or otherwise) by presenting non-facts as facts” (Thompson, 2008, 9). That is the true and the false become indistinguishable and interchangeable.

Nichols (2017) examines in more depth the effects of knowledge rejection which he terms the death of expertise, and defines as “fundamentally a rejection of science and dispassionate rationality, which are the foundations of modern civilization.” Given the fact that, at present, information is so readily available to any person with a smartphone and an internet connection, there is an endemic confusion between information and knowledge. However, as Nichols points out, the two are not at all similar, knowledge is domain-specific, it is functional and operational, it presupposes not only access to information, but also the development of specialised skills. Therefore, not everyone has knowledge in all fields, irrespective of how much information they can access. However, this is difficult to accept because it undermines people’s sense of autonomy and self-reliance, and consequently engenders feelings of rejection and hostility to institutionalized knowledge.

Strömbäck et al (2022, 1) refer to a similar trend when analysing knowledge resistance, which they define as “the tendency to resist available evidence, and more specifically empirical evidence”. Glüer & Wikforss (2022, 30) term this rejection as a form of irrationality, which seeks to negate the link between the empirical evidence and a claim or conclusion, not on the basis of other, contradictory or refuting empirical evidence, but on motivated reasoning and preexisting beliefs. Thus polarization emerges (Glüer & Wikforss, 2022, 30) and common knowledge and truth become controversial and open to debates.

What all these concepts capture is the decline of reliance on knowledge and expertise in contemporary democratic societies, a rejection of knowledge in favour of personal opinions and emotional beliefs. Truth decay subverts the very essence and promise of democratic systems which is to encourage and foster progress for all individuals.

However, all these trends, which are essentially similar, fail to capture or simply ignore the essence of the scientific method. As Keeley (1999), O’Connor and Weatherall (2019) point out, it is not that science does not make mistakes or scientists do not produce erroneous results at times. The essence is that scientists are always vigilant and scrutinize their work, through peer-review, replication, etc., openly admit when they uncover they were wrong, and constantly attempt to find ways to correct themselves and improve their research.

> Ultimately, the reason to rely on scientific knowledge when we make decisions is not that scientists form a priesthood, uttering eternal truths from the mountaintop of rationality. Rather, it is that scientists are usually in the best position to systematically gather and evaluate whatever evidence is available. The views of scientists on issues of public interest—from questions concerning the environment, to the safety and efficacy of drugs and other pharmaceuticals, to the

risks associated with new technology—have a special status not because of the authority of the people who hold them, but because the views themselves are informed by the best evidence we have. (O'Connor and Weatherall, 2019, 44)

These scientific endeavours can be thwarted not from within, as the scientists, as previously stated, have inner control mechanisms to identify mistakes and correct them, but from without, when science is manipulated to serve particular interests, or it is simply dismissed because it does not serve the policy-makers' interests. Kavanagh and Rich (2018, 26) explain that the first trend promoting truth decay, the increasing disagreement about facts and analytical interpretations of facts and data, affects not only recent research, in whose case the data may be still inconclusive or in need of further verification (e.g., a new possible cure for cancer), but also clearly established and confirmed scientific conclusions (e.g., such as the vaccines are beneficial, climate change is real). They notice that there is an "increased divergence between public attitudes and facts and data emerging from scientific research." In fact, people seem to be rejecting facts and data in favor of personal experiences, personal stories, and opinions. And this rejection fuels a vicious circle, in which people refuse to learn more about scientific findings and thus know less, and rely even more heavily on their personal interpretations, and accept opinions as facts, because opinions they can comprehend more easily than sometimes very complex scientific facts.

This rejection also favors those whose interests are to manipulate and dismiss scientific findings. A famous example given by O'Connor and Weatherall (2019) is called "the Tobacco Strategy". It was first developed by tobacco manufacturers in the 1950s, when physicians first drew the alarm with respect to tobacco-induced ailments, including lung cancer. In essence, the strategy relies on fighting science with more science. The tobacco producers could not deny the fact that tobacco caused serious diseases. However, they could induce doubt, by funding research into other causes for those types of cancer frequently associated with tobacco consumption, and by concluding that research into tobacco effects on health were not as definitive as they seemed. Not that they were wrong, but they were not definitive. This strategy was widely successful, and it led to decades of delays in health regulations regarding smoking. It was then translated successfully in other areas, such as sugar consumption. In essence, the idea is that science is not dismissed, it is merely drowned out; thus the public becomes confused by the myriad of possible causes for various ailments, and confusion leads to inaction.

Levy explains that epistemic authority is properly constituted when it has the right structure, namely a "inquirers, methods and results are publicly available (especially, but not only to other members of the network), inquirers are trained in assessing knowledge claims according to standards relevant to the discipline, and rewards are distributed according to success at validating new knowledge and at criticizing the claims of other members of the network" (Levy, 2007, 188).

He also points out that knowledge thus produced is "deeply social". There are some kinds of knowledge that once acquired are always available to individuals (e.g., the multiplication scale); however, the more specialized knowledge is only available in these networks, in which specialists from different fields contribute. Levy states that cutting ourselves from these networks of knowledge does not merely breed sceptisicm and distrust, it "more radically means cutting ourselves off from our own best epistemic techniques and resources" (Levy, 2007, 188) which can stop societal progress.

Epistemic trust is the foundation of democratic progress, but in order to reach its potential, it needs to be doubled by institutions willing to put into practice scientific discoveries, and to act in good faith with respect to the available knowledge. This leads to the second type of trust that is affected by disinformation and the contestation it entails.

Dismantling institutional trust

What disinformation does is undermine institutional trust (powerful secret actors with nefarious interests have corrupted said institutions) and epistemic trust (scientists are not objective, they do not engage in accurate research, but respond to financial stimulants). This creates a vicious circle in which disinformation weakens trust, leading to a state of uncertainty, anxiety, perceived lack of control which enhance the need to find meaning and explanations that, in turn, are readily available and easily understandable in conspiracy theories (as seen in chapter 1.3).

Democratic institutions need the citizens' trust and support, to the same extent as they need the best knowledge available to inform their policies. However, if societies are polarized, and citizens are reluctant to trust anybody except for their close circle of family, friends, acquaintances, this affects the institutions' abilities to perform their societally appointed roles. For example, if citizens do not trust that the medical system works for their welfare, then they are unwilling to fund it through their taxes, which will lead to a decline in the quality of the services it provides. If citizens do not believe that the government works in their interests, but rather that it is the slave of a global cabal whose goal is to destroy society as we know it, then they will not participate in elections, on the one hand, and/or not obey by those elections results once they take place (e.g., the 1/6 insurrection in the USA, contesting the results of the presidential elections).

Snyder (2018) introduces the concept of strategic relativism that is at the core of Russian propaganda. Strategic relativism is defined as the transformation of "international politics into a negative-sum game, where a skillful player will lose less than everyone else." The idea behind strategic relativism is that progress, reform, must seem impossible in all societies, not only in Russia, and therefore, Russians would no longer seek it, and Europeans and Americans would no longer see it as viable, and their democratic systems would therefore disintegrate.

> If the citizens of Europe and the United States joined in the general distrust of one another and their institutions, then Europe and America could be expected to disintegrate. Journalists cannot function amidst total skepticism; civil societies wane when citizens cannot count on one another; the rule of law depends upon the beliefs that people will follow law without its being enforced and that enforcement when it comes will be impartial. The very idea of impartiality assumes that there are truths that can be understood regardless of perspective.

Researchers notice that one of the most important pillars of participatory democracy is a strong and independent mass media ecosystem. It is the institution meant to safeguard democracy by bringing to light and holding accountable any deviations from the rule of law, from the principles and tenets of democratic societies. However, at present, the mass media is under assault as well. Several factors have led to the current fragmented media ecosystem, with highly ideologized broadcasts, and polarized audiences, in which the very foundation of a free press has been subverted, as it is no longer viewed as an objective informer, as a promoter of facts not opinions, but as just another biased voice in an already very crowded public space.

Kavanagh and Rich (2018) explain that the transformation of the media started with the emergence of the 24/7 news channels which encouraged the development of the talk shows in which experts and/or pundits appear and present their opinions on the events, rather than the events themselves, thus further blurring the line between facts and opinion which leads to truth decay. Moreover, these media conglomerates are dependent on financing and advertising, meaning that there are financial constraints they operate against in presenting the news; they have to respect the ideological lines dictated by the owners. Ratings also matter, which means that the news no longer presents events and facts objectively but slanted in such a way as to be appealing to their target audiences and thus keep them glued to the screen and to the channel. Mass media has relinquished its educational goal, and has become subjected to viewer's whims, biases, beliefs that they have to respect and reinforce.

The news thus becomes a daily spectacle in which the shows which elicit the most emotional responses, the most anger, anxiety and revolt are the most successful, to the detriment of those which present facts objectively, without ideological bias. If it lacks emotion, the news is not compelling, which means it is not believable, and the viewers switch channels in search of the sensational.

This already challenging, fragmented and emotionally charged media ecosystem, is further complicated by the emergence of social media. More and more people, all over the globe, report that they get their news from social media platforms. Snyder (2018) explains that "the internet is an attention economy, which means that profit-seeking platforms are designed to divide the attention of their users into the smallest possible units that can be exploited by advertising messages," and the news on these platforms is not tailored to encourage reflection, but to fit and decreasing attention span and the hunger for reinforcement, thus forming a "neural path between prejudice and outrage" which does not encourage action, but rather a continuous spiral of discontent and distrust.

Filter bubbles (Pariser, 2011), selection algorithms (Oremus, 2016), and echo chambers (Sunstein, 2001) have customised the content that users access, based on their beliefs, and thus have personalised the information and knowledge they gain, as well as the interactions they engage in. Sunstein identifies three main problems that this continuous filtering raises:

1. Fragmentation, the creation of communities that only engage and listen to their own members and reject any external interventions. The danger here goes beyond mere fragmentation and the erosion of public discourse; fragmentation can lead to extremism, hatred and even outright violence.
2. Information is a public good, which should be shared with others who might benefit as well. What one person learns, they share with others and thus knowledge remains the social product that Levy spoke of. However, in a system designed to insulate people from information that may contradict their already held beliefs, this becomes increasingly if at all possible, as people will relate to each other only the little they know, without accepting anything new.

3. Not understanding the connection between freedom and the relationship between consumers and citizens. When speaking of consumers, filtering is a way of getting them what they want in a short amount of time. However, citizens need more than satisfaction that their beliefs are the same as others. They also need to be exposed to contrasting or even opposed beliefs in order to truly examine a problem from all angles and come to the best-informed decision. This is what freedom is truly about in a democracy. Not the freedom to insulate oneself, but the freedom to express oneself and come into contact with others' expressions.

Social media, as an increasing insulated environment in which people interact only in small, tailored groups, lead to accentuated truth decay since facts and opinions are not differentiated and the content is actually segregated to be in tune with the groups' pre-existing beliefs which are thus reinforced (Kavanagh & Rich, 2018). Consequently, ruptures and polarization increase in societies, as people insulate themselves in communities with no contact with opposing or diverging views, where debate does not exist, only a spiral of confirmation and "tribal" belief reinforcement(Kavanagh & Rich, 2018; McIntyre, 2018). Nichols presents the results of a study performed at University College of London which revealed the fact that despite having more available sources of information than ever before, students limited their reading to the very first lines of an article and then moved to the next. He explains that this is not actually reading, but scrolling in search of confirmatory details for a pre-existing belief, and marks the unwillingness to engage in the attempt to follow and understand contradictory articles. Ultimately, this disengagement is detrimental to democracies because it "undermines role of knowledge and expertise in a modern society and corrodes the basic ability of people to get along with each other in a democracy" (Nichols, 2017). The factual common ground so necessary for informed democratic debates in the public sphere with respect to how and what societies should do is fractured, due to lack of adherence to common facts and consensual truth, and to lack of constructive debates. If one compounds truth decay and trust decay, one has the perfect fertile ground for conspiracy theories and unfounded rumours to disperse in society.

Interpersonal trust as a vector for spreading disinformation

It is important to notice that interpersonal trust remains intact. Its foundation is mainly emotional, therefore contestations of facts, scientific truth and explanations do not weaken it. As Kavanagh & Rich expound, "social relationships and networks play a large role in the formation of beliefs and attitudes" (2018), however, they severely limit the diversity of information that one comes in contact with and reinforce echo chambers in which information is never externally verified and confirmation of even the most outrageous belief is readily available. In search of personalised content, people have personalised knowledge and facts, and while people are entitled to their own opinions, they are not entitled to their own facts. Unfounded rumours and conspiracy theories circulate freely in close(d) communities, in which people share them with their peers, who accept them on the basis of interpersonal trust.

But there are inherent limitations that come with relying too extensively on individuals and their knowledge and understanding of the world. Individuals do not possess as vast and as detailed model of the physical and social world as they come to believe. Levy (2007, 184) explains that these models are actually located outside of individual cognition, in a social network; they are in fact external representations, which require fewer individual resources. This means, that when an individual has a problem, they do not have to find the solution on their own. They know where to go, who to contact, to provide them with the best approach. But this requires trust, and trust outside one's own internal knowledge, or close(d) group knowledge. This brings back the issue discussed in 4.2, regarding epistemic trust, trust that there are knowledge communities for various fields whose expertise an individual should take advantage of and have confidence in and in 4.3 regarding institutional trust, trust that the democratic institutions have their best interests in mind when promoting certain policies to aid with their concerns and problems.

In Muirhead and Resenblum's opinion, this is a matter of common sense, defined as "our acceptance of the intractable facts about the world and our already existing shared experience and understanding about our social world" (Muirhead & Resenblum, 2019, 127). Common sense, they argue, comprises "shared perceptions, experiences, and moral sensibilities, which make democracy possible." (Muirhead & Resenblum, 2019, 128) "Common sense creates a world in which it is possible for people to exchange reasons and feelings that "make sense" to one another—even under conditions of diversity and political conflict. Common sense is a resource against the tyranny that imposes its own reality." (Muirhead & Resenblum, 2019, 128) But they see common sense as affected and even betrayed by conspiracism (see 1.3), which contests facticity and common interpretation. In the absence of these two factors, political discussions become impossible as their no common, shared ground of understanding for them to rely on and resort to.

The emphasis on individuals is seen as detrimental for democracies which presuppose communities by Banti (2020, 47) as well. His grim view is that once individual success is the only yardstick for measuring welfare, then the collective nature of the democratic societies is discarded and the individuals can become mere anonymous pawns that follow the light of the successful individuals, without adhering to the benefits that cooperation and collaboration could bring about.

Snyder also analyses another effect that promoting individuality and emotions to the detriment of factuality has on democratic societies. His analysis highlights the Russian propaganda playbook and its goal to discredit democratic societies, not only in the eyes of the Russian society, but also in the eyes of citizens from democratic societies. If citizens doubt everything and trust nothing, then they cannot have sensible debates about reforms and progress and cannot trust each other to organize politically to change the status quo. The discreditation of knowledge paves the way for inaction, for turning emotion into the only sterile response to any discontent.

There is another downside to people trusting only like-minded individuals and grouping themselves according to their beliefs rather than remaining open to debate and exposed to new information. In an experiment carried out by Schkade, Sunstein and Hastie (2007), they discovered that if people with similar views were grouped together and asked to deliberate on certain issues that were ideologically laden, even a 15-minute debate led the most moderate participants to adopt the most extreme views in the group: "deliberation increased consensus and decreased diversity". Once the deliberation was over, the groups were more extreme in their convictions and fewer people remained in the middle. The researchers identified four contributing factors to increased polarization:

1. Informational influences – the group members provided information that pertained to one extreme of the ideological spectrum. As that was the only information provided, there seemed to be consensus and the group became more radical and uniform.
2. Corroboration effects – when people do not have confidence that they know enough they stay away from extreme viewpoints and have a more neutral position. However, agreement from others, who corroborate what they might have only tentatively thought, encourages them towards the extreme point of that belief. People become more confident once they find out that others share their beliefs.
3. Social comparison – people share views because they might want to be perceived favorably, by the other members of the group, and will adjust their positions to be more in line with those of the group.
4. Shared identity – people want to feel they belong, they are part of a group, they have a shared identity, and they are willing to polarize in order to achieve this sense of shared identity, to be part of the ingroup and more clearly separated from the outgroup.

The problems arise as this phenomenon of division into small groups who have similar beliefs and reinforce one another to the point of extremism is deeply facilitated by online platform that allow people to pick and choose who they listen to, and by the increasing tendency of mainstream media to conform to people's preexisting beliefs, rather than try to educate and mould them. As Kavanagh and Rich (2018, xvi) explain, the effects on democratic societies are dire: civil discourse is eroded as the parts involved are unwilling to listen to each other; political paralysis appears as there is increasing uncertainty about facts which creates difficulties in agreeing what the best policies would be; citizens feel that the government is further letting them down by not acting which leads to disengagement from political action and civic institutions; and in general policy becomes uncertain and even inoperable at a national level.

Given all these challenges, it is more important than ever to try to identify means of countering them and of rebuilding trust.

Main challenges and means to overcome them

In light of the aspect presented above, the greatest challenges in tackling the democracy stifling effects of disinformation require a multi-faceted approach. A survey of the literature has revealed several methods that could be put into practice and that, if adhered to and carried through, could help rebuild trust in epistemic authorities, in democratic institutions and processes, and, in democracy itself.

1. **Believe in truth.** Snyder explains that fighting tyranny and upholding democratic systems starts from the individual and from individual action. "To abandon facts is to abandon freedom. If nothing is true, then no one can criticize power, because there is no basis upon which to do so. If nothing is true, then all is spectacle. The biggest wallet pays for the most blinding lights" (2017, 65). Speaking and asserting truth helps spread the message and uphold the criteria of judgement for democratic institutions and experts.

2. **Speaking truth to disinformation** (Muirhead & Rosenblum 2019, 14, 116-120)[35]. This starts with small steps that presuppose a change in the way in which citizens interact with one another. It is the shift from small communities of shared and similar beliefs who reinforce one another's opinions to a watchful and engaged civil society that bears witness and upholds directly and transparently established and proven knowledge. This step combined with the mass media reassuming its role as a purveyor of information, facts, data, not (only) opinions, could lead to countering the societally corrosive effects of disinformation.
3. **Enacting democracy**. This refers to "the scrupulous and explicit adherence to the regular forms and processes of public decision-making." It entails a consistent, deliberate and sustained response to disinformation in order to mitigate the increasing distrust in democratic systems. "Enacting democracy makes government legible. That is, it gives citizens reasons to understand and appreciate the meaning and value of institutional integrity and ordinary democratic processes" (Muirhead & Rosenblum, 2019). This process, if repeated with perseverance, can lead to a cumulative effect which can help relegitimize democratic processes and institutions and rebuild institutional trust. It also means that the processes themselves are openly articulated, presented in a transparent manner to the citizens, thus educating them on how the institutions function and encouraging them to take an active part. Citizens need to witness institutional integrity and transparency in order to regain trust.
4. **Defending knowledge producing institutions**. These institutions and the scientists and experts that serve them, might be wrong at times, either due to errors or intentional corruption. However, scientific endeavours are continuously monitored and evaluated and their mistakes are more likely to be discovered and corrected. The experts and policy-makers should be held accountable for their judgements; but they should not be dismissed and distrusted. Rather, they should be questioned openly, they should be required to present their evidence, explain their reasonings, have their conclusions reviewed. The world of expertise and democratic policy-making institutions should be bridged so that they could inform each other and ensure societal progress.

   Sunstein (2017) also proposes two methods that could help rebuild trust and societal cohesion:

5. **Rebuilding the media ecosystem** by supporting "general-interest intermediaries" (Sunstein 2017, 26-27) such as newspapers, magazines, broadcasters, that promote facts presented as facts and opinions presented as opinions and thus encourage debates on the same common, shared framework of knowledge which allow people with diverse opinions to interact with one another and to be exposed to various points of view, without losing sight of the facts. "A system in which individuals lack control over the particular content that they see has a great deal in common with a public street, where you might encounter

[35] They proposed them to tackle new conspiracism (see 1.3 for definition), but they can be adapted to counter the effects of disinformation as well.

not only friends but also a heterogeneous array of people engaged in a wide array of activities (including perhaps bank presidents, political protesters, and panhandlers).” This helps create a democratic system that fosters deliberation not among like-minded individuals but among all citizens. Snyder also speaks of the importance of investing time and effort into the mainstream, objective mass media outlets. He argues that it is every citizen’s obligation in democratic societies to make a custom if investigating and not taking things for granted, just because they are pleasing to hear or easy to understand. “Figure things out for yourself. Spend more time with long articles. Subsidize investigative journalism by subscribing to print media. Realize that some of what is on the internet is there to harm you. Learn about sites that investigate propaganda campaigns (some of which come from abroad). Take responsibility for what you communicate to others. It is not only the responsibility of epistemic and institutional authorities to promote transparency and information, it is just as much the responsibility of each citizen to stay engaged and active in the sometimes tiring process of countering disinformation and maintaining a healthy ecosystem.

6. **Participating in public forums.** Sunstein (2017, 40-44) promotes the idea of forming public forums, on the basis of the freedom of speech in public places, with three main goals in mind:

   a) To give speakers access to a range of citizens with varying opinions and beliefs and thus expose those citizens as well as the speakers to new information and other points of view that may make them reconsider their positions and/or may bring to light points of public discontent, thus aiding the improvement of public policies and institutions.

   b) To give access not only to heterogeneous groups of citizens but also to specific groups and/or institutions that citizens may be discontented with so they could make their dissatisfaction and complaints known.

   c) To enhance the possibility that people will come into contact with a wide variety of people and views, they will have unexpected encounters that expose them to diverse viewpoints and experiences, thus reducing risks of polarization and promoting and (re)developing civil discourse and empathy.

A similar idea is proposed by Snyder as well when he argues in favor of practicing “corporeal politics. Power wants your body softening in your chair and your emotions dissipating on the screen. Get outside. Put your body in unfamiliar places with unfamiliar people. Make new friends and march with them.” (2017, 83) Stepping out of the boundaries of what is known and familiar and experiencing new ideas, leads to a less fragmented society, in which individuals are more open to debate and free exchange of ideas in a constructive manner, in the public sphere, that could lead to the improvement of the democratic societies they live in.

References:

1. Banti, Alberto Mario. *La democrazia dei followers. Neoliberalismo e cultura di massa*. GLF: Editori Laterza, 2020.

2. Glüer, Kathrin & Åsa Wikforss. “What is knowledge resistance” (29-48) Knowledge Resistance in High-Choice Information Environments. Strömbäck, J., Wikforss, Å., Glüer, K., Lindholm, T., & Oscarsson, H. (eds.). New York & London: Routledge, 2022.
3. Hardin, Russell. “Conceptions and Explanations of Trust” (3-39) in Cook, Karen, ed. Trust in society. New York: Russell Sage Foundation, 2001.
4. Kavanagh, Jennifer & Rich, Michael D. *Truth Decay. RAND report*. Santa Monica, California: the RAND Corporation, 2018.
5. Larson, H. J., Clarke, R. M., Jarrett, C., Eckersberger, E., Levine, Z., Schulz, W. S., & Paterson, “Measuring trust in vaccination: A systematic review”. Human vaccines & immunotherapeutics, 2018, 14.7: 1599-1609.
6. Levy, Neil. “Radically socialized knowledge and conspiracy theories”. Episteme, 2007, 4.2: 181-192.
7. McIntyre, Lee. Post-truth. MIT Press, 2018.
8. Moore, Alfred “On the democratic problem of conspiracy theories” (111-134) in *Conspiracy Theories and the People Who Believe Them*, ed. Joseph E. Uscinski. Oxford University Press, 2019.
9. Möllering, Guido. The nature of trust: From Georg Simmel to a theory of expectation, interpretation and suspension. Sociology, 2001, 35.2: 403-420.
10. Muirhead, Russell & Nancy L. Rosenblum. *A Lot of People Are Saying The New Conspiracism and the Assault on Democracy*. Princeton and Oxford: Princeton University Press, 2019.
11. Nichols, Tom. *The death of expertise: The campaign against established knowledge and why it matters*. Oxford University Press, 2017.
12. Nicodemo, Francesco. *Disinformazia. La comunicazione al tempo dei social media*. Venezia: Marsilio Editori, 2017.
13. O’Connor, Caitlin and James Owen Weatherall. *The Misinformation Age. How False Beliefs Spread*. New Haven & London: Yale University Press, 2019.
14. Oremus, Will. Who Controls Your Facebook Feed. A small team of engineers in Menlo Park. A panel of anonymous power users around the world. And, increasingly, you, 2016, available at http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html?via=gdpr-consent
15. Pariser, Eli. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.
16. Schkade, D., Sunstein, C. R., & Hastie, R. “What happened on deliberation day”. Calif. L. Rev., 95, 915, 2007.
17. Snyder, Timothy. *On Tyranny. Twenty Lessons from the Twentieth Century*. New York: Tim Duggan Books, 2017.
18. Snyder, Timothy. *The Road to Unfreedom: Russia, Europe, America*. Tim Duggan Books, 2018.
19. Strömbäck, Jesper, Åsa Wikforss, Kathrin Glüer, Torun Lindholm and Henrik Oscarsson. “Introduction. Toward understanding Knowledge Resistance in High-Choice Information Environments” (1-28) in *Knowledge Resistance in High-Choice Information Environments*. Strömbäck, J., Wikforss, Å., Glüer, K., Lindholm, T., & Oscarsson, H. (eds.). New York & London: Routledge, 2022.
20. Sunstein, Cass R. *#republic. Divided Democracy in the Age of Social Media*. Princeton and Oxford: Princeton University Press, 2017.
21. Sunstein, Cass R. *Echo Chambers: Bush v. Gore, Impeachment, and Beyond*. Princeton, N.J.: Princeton University Press, 2001.
22. Thompson, Damian. *Counterknowledge. How we surrendered to conspiracy theories, quack medicine, bogus science and fake history*. New York, London: W. W. Norton & Company, 2008.

2.4. Technological factors: social media, deepfakes, bots, swarms of bots

Cristina Arribas Mato, Manuel Gertrudix, Rubén Arcos

Abstract

The section delves into the knowledge of technologies that, such as deepfakes, bots and swarms of bots, are used to amplify the effects of disinformation on social networks. It explains how these technologies operate and the effects they have on the propagation of disinformation, but also on misinformation and the manipulation of information. The different types of bots in social media are defined and classified and the function of each type is detailed. From this, some strategies aimed at combating its effects are indicated, including the automatic detection of these systems. In addition, it explains how the deepfakes and forgery's function, and how to counter it, and examples of online applications are given that allow them to be generated easily. Finally, given the concern about the malicious use of deepfakes, it analyses of automatic detection methods.

Main research questions addressed

- What role does technology play in the spread of misinformation and the manipulation of information?
- How bots, chatbots and trolls have been using in information influence campaigns?
- What strategies can be used to combat the effects of bots and automatic detection systems?
- How function the deepfakes and forgeries, and how counter it?

Information influence operations by hostile actors in social media

Social media are one of the most obvious manifestations of the technological revolution. They have contributed decisively to the current paradigm shift in communication, leading to the loss of the information monopoly held by traditional media, these are newspapers, press and television, and giving way to a kind of model of journalism 2.0, where each citizen stands as a potential producer and disseminator of content. Compared to the unidirectional traditional information model, where the consumer assumes the exclusive role of receiver, a bidirectional and plural model is now established. Within this model, social networks become the preferred channel of information consumption. This situation led to a darkening of the information activity. The intrinsic characteristics of social networks and instant messaging platforms, as well as the preferential use of smartphones to access the Internet, opened the door to information influence operations by hostile actors.

The immediacy in the distribution of content, together with very high production, both made possible by social networks, increase the difficulties of consumers to discern the veracity, credibility, and reliability of these, as well as the intentionality of the actors who manufacture and disseminate the information.

In this context, the demand for responsible consumption of information by the public receiving the messages is outlined, but also the involvement of social media platforms to combat the dissemination of disinformation. On the part of the public, this responsible consumption should be based on a critical approach both to the contents and its verification, but also to the analysis of arguments, and the accounts from which they are distributed. Technology platforms must participate by identifying disinformation content, but also by preemptively detecting those accounts that may be potentially malicious.

The role of AI in this context is central, and can also be attributed to an ambivalent character, supporting the generation of content that serves hostile actors such as Deep fakes, or its dissemination through bots or chatbots, the selection of content according to its prediction of viralization, but it also plays an essential role in combating disinformation and its effects, detecting content, as well as potentially malicious social media accounts.

Use of bots, chatbots and trolls in information influence campaigns

The use of social media bots (SMB) together with the so-called trolls as amplifiers of informative influence campaigns in different contexts (political, health, climate change, etc.) has been the object of attention by researchers in the field of communication and computer science (Ferrara et al., 2016; Varol et al., 2017). Social media bots (SMBs) can be defined as automated or semi-automated accounts that mimic human behaviour and interact with other accounts, (Abokhodair et al., 2015). These accounts generally belong to a coordinated net (botnet) and are directed by a Master Bot.

On the other hand, trolls are accounts managed by a user who serves the interests of a specific actor. The interaction between troll accounts and bots in informational influence and disinformation actions and campaigns is widely documented (Broniatowski et al., 2020; Badawy et al., 2018).

Among the events that have been the focus of researchers involving large-scale social media bots are the 2016 US presidential campaign (Luceri et al., 2020; Badawy et. al, 2018; 2019; Bessy & Ferrara, 2016), the anti-vaccine debates of the period 2014-2018 (Broniatowsaki et al. 2018), the French presidential elections of 2017 (Ferrara, 2017), the Catalan referendum for independence (2017) (Stella et al., 2018), Brexit (Bastos and Mercea, 2018) or the Covid-19 infodemic (Uyheng et al., 2020).

Subrahmanian et al. have shown that social bots represent an important percentage of the total number of the Twitter social accounts. Between a 5 and 9%, according a 2016 study (2016), which produced the 24% of the total tweets (Morstatter et al., 2016), in turn, a 2022 research conducted by the Israeli cybersecurity company CHEQ showed that from the 5.21 million website visits analysed that came from Twitter, a 11.71% were from bot accounts.[36]

According to their characteristics, there is a wide spectrum of social media bots, from those that automate simple tasks (e.g., sharing or liking), other hybrids directed by humans but that have automated tasks to acting agents equipped with artificial intelligence (Assenmacher et al., 2020) and include machine learning (ML) and Natural Language Processing (NPL) algorithms. Automated accounts that participate in informational influence or disinformation campaigns only constitute a limited percentage of the total set of bots operating on social networks, nor are all targeted for malicious purposes (different types of spam, abuse of credentials, scam, fraud, pornography, e.g.,) (Adewole et. al., 2016). Orabi et al., (2020) applied the following classification for bots in social media:

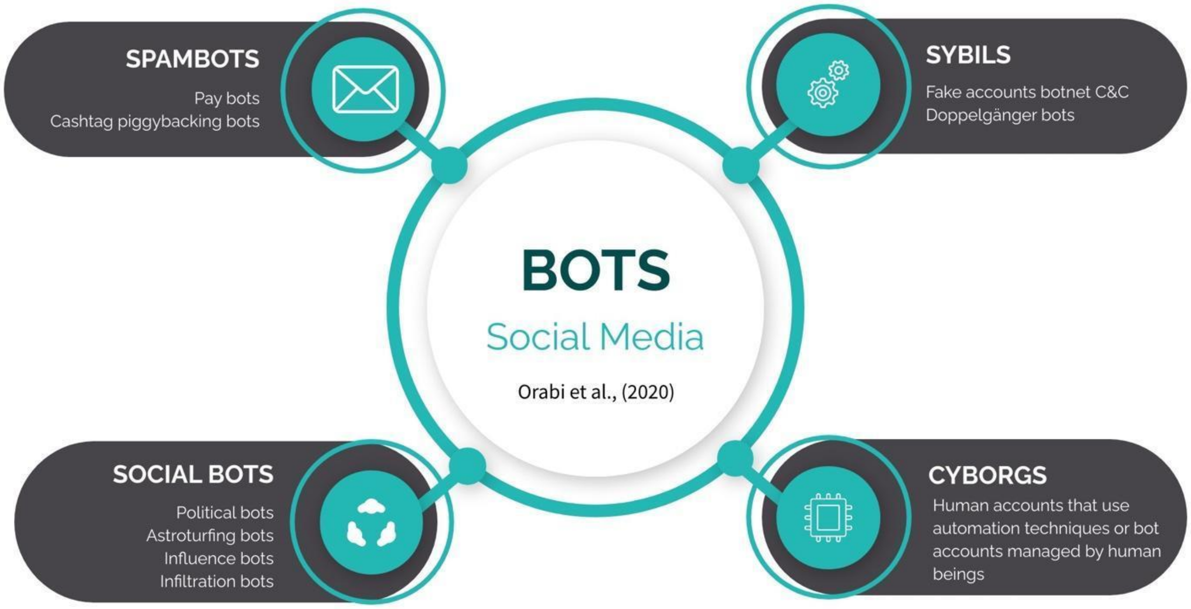


Figure 1. Bots in social media. Source: Own elaboration from Orabi et al. (2020)

- **Spambots:** Bots that distribute malicious URLs, unsolicited messages and hijack trending topics contaminating the conversation with other content (Ibídem, p. 5)
    - Pay bots
    - Cashtag piggybacking bots[37]
- **Social bots:** "Computer programs designed to use social networks by simulating how humans communicate and interact with each other" (Ibídem, p.5)
    - Political bots

[36] Johansen, Alyson Grace. "What's a Twitter bot and how to spot one." Norton Emergency Trends, September 5, 2022.

[37] Coordinated net of bots that promote low-value stocks by exploiting the high-value ones (Cresci et al., 2019)

  - Astroturfing bots
  - Influence bots
  - Infiltration bots
- **Sybils:** “Pseudonymous identities, i.e., user accounts, used for a disproportionately large influence” (Ibídem, p. 5)
  - Fake accounts used for botnet C&C
  - Doppelgänger bots[38]
- **Cyborgs:** “Human accounts that use automation techniques or bot accounts managed by human beings” (Ibídem, p. 5)

Also noteworthy are chatbots, which allow interaction with users, and especially the so-called Conversational AI, -state-of-the-art chatbots - that also incorporate Natural language understanding (NLU), Machine learning, Deep learning, and predictive analytics which allows them to make decisions, imitating human behaviour (Nuacem AI, 2021). A good example is ChatGPT developed y Open AI and based on GPT-3 algorithm, which understands human discourse, is multilingual, doubt premises, refuse inappropriate offers, and can create texts following instructions such as poems, legal texts, games, etc. Its potential to produce disinformation, reducing the work necessary to write it, while expanding its reach and effectiveness by focusing on specific targets, has been reported by experts from Georgetown University’s Center for Security and Emerging Technology. Another example is Replika a chatbot whose aim is to become the best friend of the users learning from their inputs. The application can lead to a self-isolation of the individual and the reaffirmation of one's own thought even if it is consistent with violence, hate speech or conspiracy theories.

The risks associated with these technologies related to the automation of tasks and their mimicry with human behaviour are also affected by different factors that hinder the investigation of information influence campaigns, as well as the effectiveness of the countermeasures launched. These factors include the confluence of domestic and foreign actors in disinformation campaigns, the offer of influence services through bots by private companies, as well as the "conflict over distinguishing harmful disinformation and protected speech" (Sedova et al., 2021). The difficulties derived from the use of proxies, like local actors, generally at the service of the interests of foreign state actors. Disinformation campaigns, together with the use of social bots and the tendency to work on different social media platforms to reach wider audiences, make it extremely difficult to trace the origin source of the information and to be able to establish the attribution of authorships.

On the other hand, the democratization of disinformation thanks to companies that provide influence services based on the use of bots, has led to its identification being essential, being an important niche for researchers and companies as a line of business given the rapid evolution of AI technologies and their use by hostile actors.

Strategies to combat the effects of bots and automatic detection systems

[38] Cloned Twitter profiles to use them in malicious activities (Goga et. al., 2015)

Tackling disinformation from preventive approaches is the main challenge to combat it. The development of newly created social media bot account detection systems, even before they start posting, is essential in identifying disinformation campaigns in their latent phase (Arcos & Arribas 2023).

Different approaches have been valued and implemented to combat the effects of social media bots. Orabi et al. (2020), conducted a systematic literature analysis on strategies for bot detection in social media. The authors also collect some references to works in which other ways of combating them are addressed. Thus, Almerekhi & Elsayed (2015) are oriented more toward the automatic classification of posts rather than to the detection of accounts. Other authors focus on identifying especially vulnerable groups or groups in order to develop defensive strategies (Halawa et al., 2016).

Regarding strategies for the detection of bots, researchers and analysts have explored different approaches. A first trying at classification was carried out by Ferrara et al. (2016) and distinguished between graph-based, feature-based, crowdsourcing-based and combined approaches. Later, the taxonomy was widened with additional subcategories by Adewole et al., (2017).

Van Der Walt and Eloff (2018) also contributed to summarize the different approaches found in the literature on the detection of bots. They identified different key features to explore fake content linked to the social media account; analysis of the profiles; activity features such as the timeframe between account creation and first post; the time between posts; stance and sentiment towards topic issues, and relationship with similar accounts or target accounts.

From all previous classifications, Orabi et al. (2020) propose a taxonomy guided by the methods employed:

- Graph-based
- Machine learning (ML)
    - supervised: based content, based behaviour
    - semi supervised
    - unsupervised: based content, based behaviour
- Crowdsourcing based
- Anomaly based: action-based, interaction-based

Following the Orabi et al., taxonomy together with the later contribution by Ilias & Roussaki (2020), the below models are identified.

<u>Graph based:</u> A graph G is a set of nodes V (G) in a plane and a set of lines (links) E (G) of a curve, each of which either joints two points or joins to itself (West, 1996). These structures allow to representation of the relationships that occur between the different accounts of social media, being possible to identify the botnets. Among the works based on this model, we found (Boshmaf et al., 2016), focused on the detection of profiles more vulnerable to the influence of these bot accounts stand out. The system is based on the attribution of higher weights to those accounts that are real compared to fake.

Machine Learning (ML): Defined as a field of study that allows computers to learn without the need to be programmed. This learning, in the case of bot detection, can be supervised (a classifier learns how to identify accounts, based on labels), semi-supervised (it uses partially labelled data) or unsupervised (the algorithm by itself clusters the input data, labels are not necessary). Within these types, two subcategories have been identified: based behaviour and based content.

An example of supervised learning based on behaviour is BotorNoT (Davis, Varol, Ferrara, Flammini, & Menczer, 2016) a system that determines -from more than one thousand different features - the likelihood that a given account is a bot. It is available for public use through a website.

Cresci et. al., (2016, 2017) developed the digital DNA technique to analyze the collective behavior of social network users to detect botnets. This technique is based on "encoding the behaviour of an account as a sequence of bases, represented using a predefined set of the alphabet of finite cardinality consisting of letters such as A, C, G and T" (Orabi, p. 8). Within this set also can be highlighted the Convolutional Neural Networks (CNNs) detect spam on Twitter both at the account and tweet level (Ilsa & Roussaki, 2020).

Among the unsupervised systems based on the identification of bots based on content, we find models that are based on the search for groups of accounts that distribute the same URL (malicious), or accounts that hijack hashtags such as the DeBot tool (Chavoshi, Hamooni, Mueen, 2016b). Also, noteworthy in this group is the BotCamp tool developed by Abu-El-Rub & Mueen (2019) oriented to political conversation, which uses Debot combined with graph-based methods to model topographical information and cluster the collected bots, and a supervised model to classify user's interactions (agreement or disagreement with a certain sentiment) (Orabi, 12).

As semi-supervised systems can be highlighted clickstream sequenced and semi structured clustering (Orabi, 12)

Crowdsourcing: Includes manual identification of botnets and collect labelled datasets.

Anomaly-based: Models that detect anomalies in the interaction and activity of accounts based on the assumption that legitimate OSN users would have no motive to engage in some odd behaviours, as it will not reward them in any aspect.

Inspiring by the previous works above-mentioned, Ilias et al., (2021) conduct an experiment applying NLP and deep neural networks on 70 different features (e.g., the time between posts and retweets, length of username and profiles, following rate, uppercase and elongated word rate, tweets per day, reputation ratio, number of different sources used, the sentiment of the contents, the time between replies, age of the account...) to different datasets. The interest of this work was to determine the best selection of features through a logistic regression model. These features included: size after compression type, maximum number or mentions in a tweet, size after compression content, unique words unigrams per tweet, retweets (tweet), unique mentions ratio, the distance between username-screen name, average mention per tweet, number of mentions per word, the maximum number of URLs in a tweet...

Deepfakes and forgeries

Deepfakes and forgeries constitutes a challenge for homeland security due these are being used for malicious purposes by hostile actors. Adversary states and motivated political individuals can be served by these technologies to erode public trust in institutions and democracy, realizing fake images or videos where public figures make inappropriate comments or behaviours or share disinformation.

The effects triggered by deepfakes are highly concerning because of their realistic results, are rapidly created, and are cheap due to the freely available software and the use of cloud computing to get processing power.

Although the manipulation of images is distant in time, already in Stalin's time those members of the *nomenklatura* who had fallen out of favour were erased from the photographic archives, and the current deepfakes models date from 2017. Soon, its potential for manipulation in different contexts was noticed. In 2018, the Center for New American Security (CNAS) published a report warning about the risks of deepfakes for political manipulation and established a five-year deadline for its improvement, at that time it would be impossible to differentiate real images from deepfakes. In early 2022 an experiment conducted by Nightingale and Farid in which 315 volunteers were asked to discriminate fake face images from those that looked real, concluded that humans are not able to identify deepfakes, offering the results achieved an average accuracy of 48.2%. In addition, the authors asked participants to score the images according to the confidence they conveyed to them. The results were disturbing, as artificially generated faces appeared more reliable than real faces because they tended to look more like average faces which also generate more confidence.

Examples of employment of deep fakes can be found in the Ukraine War. Both Ukranian and Russian sides released videos compromising Putin and Zelensky with false declarations.

How do deepfakes work on?

Deepfakes (stemming from “deep learning” and “fake”), a term that first emerged in 2017, describe the realistic photos, audio, video, and other forgeries generated with artificial intelligence (AI) technologies.

Tolosana et al., (2020) categorized deepfakes (audio non-included) into the following categories according to their format and kind of manipulation:

- Entire Face Synthesis: It refers to entire non-existence images artificially created.
- Identity Swap: Consists of replacing the face of a person in a video
- Attribute Manipulation: It refers to the manipulation of a certain attribute into images (e.g., age, hair or eyes colour, gender…).
- Expression Swap: this kind includes those images or videos where facial expression has been modified.

Nguyen et al., (2022) used two different categories referred to manipulated videos according to AI algorithms:

- Lip-sync deepfakes refer to videos that are modified to make the mouth movements consistent with an audio recording.
- Puppet-master deepfakes category refers to videos of a target person (puppet), who is animated following the facial expression and movements of another person (master). (p.1)

Deepfakes are based on Machine Learning models, especially in GANs (Generative Adversarial Networks) which work with two different neural network models that are trained in competition with each other. The first one, the generator, is tasked with creating counterfeit data (photos, audio recordings, or video) that replicate the properties of the original data set. The second network, or the discriminator, is trained to identify the counterfeit data. From the results of each iteration between both networks, the generator adjusts to create increasingly realistic data and will go on -thousands or millions of iterations—until the discriminator can no longer distinguish between real and counterfeit data.

Some examples of software that generate deepfakes are:

- Which face is real? A GAN-based tool to generate entire non-existence faces.
- FaceSwap: A free tool that used autoencoder-pairing structure a kind of deep learning model.
- Deepfacelab: This open-source software is the leading solution as deepfake generator. To obtain the maximum advantage and get more realistic results, the software requires the use of Adobe after Effects and Davinci Resolve. Beyond the changing of faces, age lifting and lips manipulations functions are supported.
- Xpression Camera: based on voice2face technology. Let to launch a realistic avatar in real time that copy users´ facial expressions. It can be used in streaming services such as Twitch, or in videoconferences.
- Faceswap: open-source software to generate deepfakes from an imported video.
- Omniverse Audio2Face: Developed by NVIDIA, this tool can animate a face (predetermined) from any audio path.
- Uberduck.ia: A very powerful audio deepfake tool that let to select between thousands of voices, such as celebrities, TV characters, YouTubers, TikTokers, etc., and insert a text, upload an audio file or record audio. There is a free version.

Deepfakes detection models

Concern about the malicious use of deepfakes has become a topic of interest for governments and organizations and research on the development of automatic detection methods is on the rise, being currently an important field for computer science researchers.

Following the surveys carried out by Tolosana et al., 2020 and Nguyen et al., 2022, we can highlight the following methods (cataloged according to the taxonomy proposed by Tolosana et al., 2020):

Photographic images: Created in their entirety, different proposals have been made, including the analysis of features of the GAN-pipeline such as the colour (McCloskey and Albright, 2018), steganalysis methods such as the pixel co-occurrence matrices (Nataraj et al., 2019), or the detection of fingerprints inserted by GAN architectures using pure deep learning methods such as convolutional traces (Guarnera et al., 2020).

Swap faces deepfakes: The following approaches for identification have been developed: detection of inconsistencies between lip movements and audio speech, features related to measures like signal-to-noise ratio, specularity, blurriness, etc; eye colour; missing reflections, and missing details in the eye and teeth areas. Analysis of facial expressions and head movements; eye blinking, blood flow or combined systems based on both facial expressions and head movements have also been proposed in the literature.

Attribution manipulation: Some authors propose to analyse the internal GAN pipeline in order to detect different artifacts between real and fake images. Many studies have also focused on pure deep learning methods, either feeding the networks with face patches or with the complete face. These systems provide results close to 100% accuracy due to the GAN fingerprint information present in fake images that are used by these systems. However, recent proposals found in the literature to remove GAN fingerprints from the fake images while keeping a very realistic appearance represent a challenge.

Expression swap detection based on deep learning approaches consider both spatial and motion information (3DNN, I3D and 3DResnet approaches); steganalysis and mesoscopic, visual effects; image and temporal information through recurrent convolutional network approaches or optical flow fields to exploit possible inter-frame dissimilarities due to fake images have unnatural optical flow due to unusual movement of lips, eyes...

Governments such as the US have focused efforts on the development of detection architecture. The Defence Advanced Research Project Agency (DARPA) has developed two systems: Media Forensics (MediaFor) y Semantic Forensics (SemaFor). Recently, Facebook Inc. with Microsoft Corp and the Partnership on AI coalition have launched the Deepfake Detection Challenge to improve the research and development in detecting and preventing deepfakes.

Despite all these efforts aimed at detection, there are important limitations, and, in practice, they take us to the field of cat-and-mouse games between hostile actors trying to overcome detection systems and detection architectural designers. While the aforementioned approaches can be effective, this accuracy is based on laboratory studies where the videos are discriminated from well-known datasets of deepfakes. There are therefore no guarantees of similar performance on new materials. Studies into detection evasion show that even simple modifications can drastically reduce the reliability of a detector. On the other hand, it should also be noted that in practice the deepfake videos that are uploaded to social networks are compressed, so the image quality would make detection difficult. Other preventive approaches that are being explored are adversarial attacks on deepfake algorithms, and the implementation of blockchain systems or distributed ledger technology (DLT) to leave a record of original audiovisual materials or watermarks within the audiovisual files.

However, it also warns of the vulnerabilities of these approaches such as attacks on the integrity of the DLT itself, or the dependence on technicians and organizations that will be responsible for operating the system, or the existence of a link between the registry and the recipient of the information. Another of the formulas that are presented as effective is the awareness of potential audiences through exposure to pre-bunking or inoculation techniques (van Huijstee et al., 2021, p. 41).

References:

1. Abokhodair, N., Yoo, D., & McDonald, D. W. (2015). Dissecting a social botnet: Growth, content and influence in Twitter. Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. ACM839–851.
2. Assenmacher, D.; Clever, L.; Frischlich, L.; Quandt, T.; Heike T., Heike; Grimme, C. (2020). “Demystifying Social Bots: On the Intelligence of Automated Social Media Actors”, Social Media + Society, July-September 2020: 1–14, /doi/10.1177/2056305120939264
3. Badawy, A.; Addawood, A.; Lerman, K.; and Ferrara, E. (2019). “Characterizing the 2016 russian ira influence campaign”, Social Network Analysis and Mining 9(1):31.
4. Badawy, A.; Ferrara, E.; and Lerman, K. 2018. “Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign”. In ASONAM, 258–265.
5. Bastos, M. and Mercea, D. (2018). “The public accountability of social platforms: lessons from a study on bots and trolls in the Brexit campaign”, Phil. Trans. R. Soc. A 376: 20180003. http://dx.doi.org/10.1098/rsta.2018.0003
6. Bessi, A., and Ferrara, E. (2016). “Social bots distort the 2016 us presidential election online discussion”, *First Monday*, 21(11)
7. Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S. C. and Dredze, M. (2018). “Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate”, American Journal of Public Health 108, (October 2018): 1378-1384.
8. Chavoshi, N., Hamooni, H., and Mueen, A. (2016). “Debot: Twitter bot detection via warped correlation”, ICDM817–822.
9. Cresci, S.; Petrocchi, M.; Spognardi, A.; Tognazzi, S. (2016). “Dna-inspired online behavioral modeling and its application to spambot detection”, IEEE Intelligent Systems, 31(5), 54-68
10. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). “Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling”, *IEEE Transactions on Dependable and Secure Computing*, *15*(4), 561-576. https://www.doi.ord./ 10.1109/TDSC.2017.2681672.
11. Cresci, S.; Petrocchi, M.; Spognardi, A. and Tognazzi, A. (2019). “On the capability of evolved spambots to evade detection via genetic engineering”, Online Social Networks and Media, 9, 1-16,, ISSN 2468-6964, https://doi.org/10.1016/j.osnem.2018.10.005.
12. Davis, C.A.; Varol, O.; Ferrara, E.; Flammini, A. and Menczer, A. (2016). “BotOrNot: A system to evaluate social bots”, In WWW companion. ACM
13. Ferrara, E. (2017). “Disinformation and social bot operations in the run up to the 2017 french presidential election”, First Monday 22(8).
14. Ferrara, E.; Varol, O.; Menczer, F.; and Flammini, A. (2016). “Detection of promoted social media campaigns”, In Tenth International AAAI Conference on Web and Social Media, 563–566.
15. Goga, O., Venkatadri, G., and Gummadi, K. P. (2015). “The doppelgänger bot attack: Exploring identity impersonation in online social networks”, Proceedings of the 2015 internet measurement conference. ACM141–153.
16. Guarnera, L.; Giudice, O. and Battiato, S. (2020). “DeepFake Detection by Analyzing Convolutional Traces,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
17. van Huijstee, M.; van Boheemen, P.; Das, D.; Nierling, L.; Jahnel, J.; Karaboga, M., and Fatun, M. (2021). “Tackling deepfakes in European policy”, Study Panel for the future of Science and Technology, European Parliamentary Research Service.
18. Johansen, A. G. (2022). “What’s a Twitter bot and how to spot one.” Norton Emergency Trends, September 5, 2022.
19. Luceri, L.; Deb, A.; Giordano, S.; and Ferrara, E. (2019). “Evolution of bot and human behaviour during elections”; *First Monday* 24(9).

20. McCloskey, S. and Albright, M. (2018). “Detecting GAN-Generated Imagery Using Color Cues,” arXiv preprint arXiv:1812.08247.
21. Morstatter F.; Wu, L.; Nazer, T.; H, Carley K. and Liu, H. (2016) “A new approach to bot detection: Striking the balance between precision and recall.” In: IEEE/ACM Conference on Advances in Social Networks Analysis and Mining (ASONAM), 533–540.
22. Nataraj, L; Mohammed, T.; Manjunath, B.; Chandrasekaran, S.; Flenner, A.; Bappy, J. and Roy-Chowdhury, A. (2019) “Detecting GAN Generated Fake Images Using Co-Occurrence Matrices,” Electronic Imaging, 5, 1–7
23. Nightingale, S. J. and Farid, H. (2022). “AI-synthesized faces are indistinguishable from realfaces and more trustworthy”, PNAS, 2019(8), e2120481119 https://doi.org/10.1073/pnas.2120481119
24. Orabi, M.; Mouheb, D.; Al Aghbari, Z.; Kamel, I. (2020). “Detection of bots in social media: A systematic review”, Information Processing & Management, 57 (4), 102250. https://doi.org/10.1016/j.ipm.2020.102250.
25. Sedova, K.; McNeill, C.; Johnson, A.; Joshi, A. and Wulkan, I. (2021). “AI and the Future of Disinformation Campaigns Part 2: A Threat Model”, CSET Policy Brief CSE https://cset.georgetown.edu/publication/truth-lies-and-automation/
26. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A. and Ortega-Garcia, J. (2020). “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection”, https://arxiv.org/pdf/2001.00179.pdf
27. Uyheng, J., Carley, K.M. (2020). “Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines”, *J Comput Soc Sc* **3**, 445–468 https://doi.org/10.1007/s42001-020-00087-4

3. MITIGATING FACTORS FOR THE DISSEMINATION OF DISINFORMATION

Introduction

The current chapter explores the most recent research into ways of countering the spread of disinformation online. It sets out with an overview of the narrative, argumentative and discursive mechanisms that make disinformation attractive to audiences and proceeds to map out the most important developments in the field of combatting disinformation. Its aim is to identify the most relevant approaches and skills that informed citizens need to have in order to develop resilience to online disinformation and to become active in stopping the spread of said disinformation.

The chapter first analyses the reasons why disinformation is so attractive to audiences, by looking into the underlying mechanisms that support it and enhance its appeal. This analysis provides a clearer understanding and a solid foundation for investigating the most recent approaches in the development of skills, competences and attitudes that allow individuals to develop resilience to disinformation.

Countering disinformation is a multi-dimensional endeavor that presupposes an extensive and multi-layered involvement on the part of each individual who accesses online sites and social media. Developing the awareness and skills needed to identify, assess and report disinformation attempts is one of the main objectives of DOMINOES project.

Moreover, the current section also examines the most relevant institutional efforts to signal disinformation attempts as well as to develop the media literacy competences of the young generation as well as of adult citizens.

Overall, the chapter summarises best practices in countering online disinformation and can be further used in developing teaching and researching competences in the field.

Digital competences addressed:

1.1 Browsing, searching and filtering data, information and digital content;
1.2 Evaluating data, information and digital content;
1.3 Managing data, information and digital content;
3.3 Copyright and licenses.

3.1 Understanding the Narrative, Argumentative and Discursive Mechanisms that Make Disinformation Attractive

Kanchi Ganatra, Aitana Radu, Ruxandra Buluc

Abstract

This section examines three types of mechanisms which make fake news attractive to readers: *narrative*, *argumentative*, and *discursive* mechanisms. In the first part, these mechanisms are described in detail, revealing the psychological and social appeal they lend to misinformation. We define and discuss the various mechanisms at play and demonstrate why they are so successful in captivating and convincing an audience. In the second half, these mechanisms will be further illustrated through the case study of a real fake news article. This section shows clear examples of narrative, argumentative, and discursive mechanisms in action, providing insight into how fake news can be identified and avoided on social media and on the Internet in general. More broadly, an understanding of these mechanisms can allow individuals to engage more critically with the claims they encounter in day-to-day life, leading to a more informed public. The scope of this section is to shed light on the mechanisms that draw people to fake news so as to better understand how to counter their influence. The section also provides an exercise of how fake news can be identified on the basis of narrative, discursive and argumentative mechanisms employed.

Main research questions addressed

- What are the narrative characteristics of fake news?
- What are the argumentative characteristics of fake news?
- What are the discursive characteristics of fake news?
- How can they be identified in a real-life situation?

Narrative Mechanisms

Research on fake news often speaks about “alternative narratives” (Brites et al, 2019) or “storytelling” (Barber 2020) or “narratology” (Ryan et al 2004). In this context, *narratives* can be understood as an essential process in the framing and distribution of fake news. However, it can sometimes be difficult to define what “narrative" means. Abbott (2002) defines narrative as relating to a story or shared understanding within a group or as representations of events, an encompassing view. Similarly, Stein and Glenn’s (1979) schema represents narratives as compositions of one or more episodes that can be chained together in various ways. Storytelling then, as a narrative mechanism, is a powerful tool. Our lives are profoundly shaped by the stories around us, narratives that instill identities, connections, values, and beliefs. Storytelling is central to education, as characters and events can make abstract principles relatable and memorable. Drama and conflict are compelling, engaging our imagination and pushing us to think about alternative futures and creative solutions.

In the wrong hands, however, storytelling can become a weapon, used to mislead, divide, and dehumanise others. Narrative appeals can offer simple solutions to complex problems, framing nuanced events as antagonistic ‘us vs them’ battles, such as in fake news stories relating to migration. Terre des homes, an NGO working on humanitarian issues has identified several such antagonistic narratives related to migrants: (1) Migrants are criminals; (2) Migrants do not need help and/or protection; and (3) All migrants want to come to Europe[39]. Fake news often manipulates in this way, using a *narrative mechanism* to immerse readers in an alternate reality. In the following sections, we can see the various tools or mechanisms that are employed in order to harness narrative engagement for fake news.

Prominence

The *Prominence-Interpretation Theory* developed by Fogg (2002) unpacks what makes fake news attractive, engaging or appear credible. It does so by breaking down the process of approaching information, focusing on the stages in which users consume (fake) news. Proponents of this theory claim that, first a user notices something (prominence), and then they interpret what they see (interpretation). For gaining successful engagement of the reader/viewer, both the processes must take place - otherwise, it results in simply scrolling over the piece of news and no assessment is performed. In light of this information, then, it is important for a news article to be ‘noticeable’ or ‘prominent’ or to ‘grab the viewer’s attention’. If it is not prominent, the process of news consumption cannot progress.

The ways in which fake news can attract consumers' attention are creative and expansive. Some of these factors which make fake news attractive or prominent include: sensationalisation, hypernarrativity, calls to curiosity (such as through click-baiting); celebrity endorsement; etc. Budak (2019), for example, conducted an in-depth study on the fake news phenomenon surrounding 2016 presidential elections in the US. She explains that while traditional news focused more on policies related to the economy, elections, women or the environment; the more popular ‘news’ found on channels like social media was fake and of a negative or hyper-partisan nature. Building on Budak’s work, Baptista and Gradim (2020) have gathered the most frequent words used in detected fake news. These include “sex”,” death”,” corrupt”,” illegal”, ”alien” or” lie”, referring to sensational or outrageous content, unlike traditional media. This further illustrates that to inspire and sustain narrative attractiveness, fake news needs to be prominent and dramatic.

Framing

Another particularly interesting tactic used in making fake news prominent and attractive is framing. Framing refers to the use of “words, images, phrases, and presentation style” to present information, in order to influence “an individual’s understanding of a given situation” (Druckman

[39] Terre des Home. “5 Fake News about Migration”. 2019. https://www.tdh.ch/en/news/5-fake-news-about-migration.

2001). Not only do we see examples of this in our everyday lives, e.g., in the form of advertising and marketing; a wide number of studies have also shown the use of framing within journalism and policy-making. Media researchers such as Tandoc and Seet (2022), for example, have provided compelling support for what Druckman (2001) calls the framing effect through their work on "Wording as Framing". Tandoc and Seet argue that although several terms can have similar denotations, these terms may evoke different meanings in the consciousness of people based on the beliefs and values individuals associate with them.

Miko- łajczak and Bilewicz (2015) have found, for example, that in the context of the abortion debate using the term "foetus" will result in higher support for abortion compared to using the term "child" because people attribute greater humanness to "child" than to the term "foetus". Similarly, when "climate change" is used instead of "global warming," people are more likely to believe the phenomenon to be true (e.g., Benjamin, Por, and Budescu 2017; Schuldt, Konrath, and Schwarz 2011).

Framing an issue in a certain way (e.g., climate change as a political issue) can activate certain information, beliefs, and cognitions that the recipient already possesses in their consciousness (Nelson, Oxley, and Clawson 1997). In this way, frame-setting is a mechanism which allows fake news narratives to interact with pre-existing beliefs, identities and priorities of those exposed to the message, i.e., their own personal narratives.

Comprehension and Attentional Focus

When the user finds a certain article or post is sufficiently prominent, they move on to reviewing or engaging with the element, something Fogg calls 'interpretation'. At this point, if the goal is to sustain the reader's interest, it is important to gain narrative engagement. To explore this aspect further, we draw on the works of cognitive behaviourists Buselle and Bilandzic (2009), who have written extensively on how and why narrative mechanisms result in engagement. They identify several dimensions of engagement which can be interpreted as representing unique but interrelated engagement processes. They argue that in order to encourage engagement with a narrative, the presented narrative should be easy to comprehend. Interestingly, however, they also claim that although the primary activity of narrative engagement is comprehension, the audience should be unaware when comprehension progresses smoothly, and become aware only when comprehension falters. They describe this as 'attentional focus' - a phenomenon where a truly engaged viewer will *not* be aware that they are *not* distracted. Essentially, when a consumer reads a piece of news that is so engaging (or so mindless) that the reader does not notice themselves drifting in and out of it - it allows for smooth narrative engagement without any barriers to ease of comprehension.

Fake news articles often exploit this mechanism of easy narrative engagement by using simple language that is easy to understand and stay engaged in. There are rarely, if any, technical words or field-specific jargon. Past research has demonstrated that, in fact, the lexicon used by fake news is more informal and simpler in detail and in technical production, not only in the title of the piece, but also throughout the text. Horne and Adali (2017) state that "[r]eal news articles are significantly longer than fake news articles", and "fake news articles use fewer technical words, smaller words, less punctuation, fewer quotes, and more lexical redundancy." In other words, fake news requires less cognitive effort and attention to process and is, therefore, attractive to readers (Horne and Adali 2017; Baptista 2020) This is not only in line with the use of heuristics; but the use of informal lexicon may also encourage the reader to engage with topics more personally or emotionally because it is written in a manner which they find familiar or even comforting.

Emotional Engagement

Emotional engagement (feeling for and with characters) is another component of narrative engagement which Buselle and Bilandzic discuss in their 2019 research. This factor appears to be specific to the emotional arousal component of narrative engagement, but not necessarily to any specific emotion. They, alongside others, hypothesise that this emotional reaction from fake news likely only represents the arousal, however, rather than the degree of these emotions. Put simply, users confronted with fake news may likely feel an emotional response to the story but may not act on these emotions until enough repeated exposure is established; or in other words until the issue becomes part of the audiences' personal narratives.

Deep Stories

To make misinformation feel engaging, personal, and intuitive, even when the content presented seems far-fetched, the most effective fake news pieces use stories that are tailored to and targeted at specific audiences. Polletta & Callahan (2017) call these targeted pieces "deep stories": narratives that reinforce what people believe describes their lives. Fake news articles may present stories that resonate with the economic or cultural anxieties of certain groups, for example; and frame their struggles in relatable (if misleading) terms. Once that narrative hook is established, the reader is more open to accepting scenarios they might otherwise consider implausible or even offensive (Polletta & Callahan, 2017). If, for example, a user is repeatedly exposed to fake news stories that fallaciously link their economic concerns with 'threats' from migrants and minorities, they might begin to internalise this prejudice, becoming more and more vulnerable to content that reinforces the story they want to believe. The impersonal complexities of economic decline are a much less satisfying narrative than a story that blames a clear 'other'. In fact, this new found explanation for their experiences may even lead to a long-lasting belief system which encourages people to ignore other points of view on the matter. Proponents of cognitive heuristics call this deliberate ignorance the 'expectancy violation heuristic,' which is a strong negative heuristic. It presupposes that if a reader comes across information which does not align with their beliefs and values, they are less likely to find it attractive or credible and may even completely ignore it.

Expectancy violation and targeted news, in turn, are further compounded by narrative mechanisms which propagate false or conspiratorial theories, claiming to offer 'special knowledge' that mainstream or expert sources wish to hide. The fact that official sources do not report the same information might be presented - in a self-fulfilling way - as further confirmation of the conspiracy. Of course, the opposite is usually true - any article that claims to reveal a shocking 'hidden truth' or tell a conspiratorial, sensational story should be viewed with particular skepticism. The reason evidence and corroborating articles cannot be found is probably because the story presented is false, not because it is a private truth available only to the chosen few.

Verisimilitude

In more subtle ways, narrative mechanisms can "muddy the line" between isolated events and larger trends (Polletta & Callahan). Fake news may begin with anecdotes or events that are true, but later present them as connected points in a larger, untrue (or unverifiable) story about the world. Fake news can misleadingly pull a 'signal from the noise', presenting patterns and narratives that are not supported by evidence.

In fact, Introne and colleagues have brought an important perspective to the emerging conversation about fake news and false narratives through their work on pseudo-knowledge (PK). Building on previous work (Introne, Iandoli, DeCook, Yildirim, & Elzeini, 2017), they describe PK as false narratives that have begun to take on the heightened status of a plausible reality within a community. Inspired by cognitive psychologist Bruner (1986), we are informed that narrative and argumentative reasoning are two separate modes of human thought - each subject to different criteria. While arguments are judged according to their veracity (i.e., whether or not they are true), narratives are judged according to their verisimilitude (i.e., whether or not they seem plausible) (Bruner, 1986). Meaning, narratives don't actually have to be true, only *plausible* or convincing. This opens up a compelling avenue for further research on fake news, conspiracies and deepfakes. As Introne et al. have indicated, so far, researchers have focused on how misinformation spreads, how to detect it, and how to reduce its credibility. Underlying this approach, however, is the assumption that misinformation and fake news carry or reinforce false narratives. Intone et al. respond to this assumption by stating that "[T]his may be the case, but our findings demonstrate that fake news is certainly not a requirement for false narratives. Rather, the Internet allows the architects of false narratives to manufacture credibility by drawing information from many credible sources" (Bruner, 1986). Put simply, falsehoods don't necessarily have to be made of complete untruths. In reality, false narratives can be (possibly even more) attractive when built upon elements of real, verifiable news just so long as they *seem plausible*.

Argumentative Mechanisms

In addition to using narrative mechanisms, fake news also appeals to readers by using *argumentative mechanisms*. In contrast with the personalised 'hook' of storytelling, argumentative mechanisms use rhetorical flourishes and misapply logic in an attempt to intellectually convince readers of a 'truth'. Without a clear understanding of how these argumentative mechanisms look and which logical fallacies are commonly employed, fake news can appear authoritative, convincing, and rational, even while making spurious claims.

Argumentative fallacies used in disinformation

Argumentative fallacies more often than not form a very prolific base for fake news in particular and disinformation more generally. There are numerous types of fallacies, however, we will present at this stage several that are more often employed in disinformation campaigns. The table below presents definitions and examples of the most common fallacies:

	Fallacy	**Definition**	**Example**
1.	**slippery slope**	claim about a series of events that will unstoppably occur and culminate in one major, negative event	Liz Wheeler, American news anchor for OAN presents an aquarium's decision not to announce the gender of a penguin. "We should ask, where does radical leftist gender ideology lead? Do liberals want human children to be genderless? If so, why? Is this based on biology? And if not, then what? What happens when human children are raised genderless? If gender is destroyed, doens't that destroy traditional gender roles? And if gender roles are destroyed, doesn't that destroy gendered relationships? And if gendered relationships are destroyed, doesn't that destroy traditional marriage? And if traditional marriage is destroyed, doesn't that destroy the family units? And if people aren't dependent on their families, then who do they depend on? That's right, the government. Which is the goal of liberals in the first place. Don't let transgender penguins fool you."[40]
2.	**ad hominem**	an attack directed against a person's character, integrity, reasons rather than the position they are holding or the arguments they are presenting	"[I]f Hillary Clinton were a man, I don't think she'd get 5 percent of the vote. The only thing she's got going is the woman's card, and the beautiful thing is, women don't like her." (Donald Trump in the 2016 election campaign)

[40] OAN: Last Week Tonight with John Oliver (HBO) - YouTube

3.	**false dichotomy**	oversimplification of a complex situation and forceful reduction to only two options, out of which only one could be correct	“I had a choice, as well: either to trust the word of a madman, or to defend the American people. Faced with that choice, I will defend America every time. “ President George W. Bush regarding the Iraq invasion to prevent Saddam Hussein from using WMD.
4.	**post hoc ergo propter hoc**	“after this, therefore because of this” a faulty causal relationship based on the idea that if something happened before something else, then the first event caused the second one	5G towers became operational and then the COVID 19 pandemic started. Therefore, the 5G towers caused the COVID 19 pandemic.
5.	**cum hoc ergo propter hoc**	“with this, therefore because of this” If two events are happening at the same time, then a causality is falsely assumed, and one is said to cause the other.	Hospitals are full of sick people. Therefore, hospitals make people sick.
6.	**attacking the strawman**	gives the impression that the argument is refuted, without engaging with the actual argument, instead replacing it with a fake argument	Chuck Todd: You sent the press secretary out there to utter a falsehood on the smallest, pettiest thing. And I don’t understand why. Kellyanne Conway: Maybe this is me as a pollster talking, Chuck, and you know data well, I don’t think you can prove those numbers one way or the other, there’s simply no way to really quantify the crowds, we all know that. You can laugh at me all you want [Chuck Todd is starting to laugh], but… Chuck Todd: I’m not laughing, but the photos are showing… Kellyanne Conway: Well, but you are, and I think it’s actually symbolic of the way we are treated by the press, the way you just laughed at me is actually symbolic of the way we’re

			represented and treated by the press. I'll just ignore it, I'm bigger than that.
7.	**red herring**	something small or inconsequential distracts attention from a relevant aspect, idea, argument	The reporter's question: Can you envision a way of supporting the universal background checks bill? Senator Lamar Alexander's answer: Video games are a bigger problem than guns because video games affect people.

Table 5. Definitions and examples for argumentative fallacies

Zompetti (2019) outlines a few different types of argumentative mechanisms. One common strategy involves distracting and deflection through pointing out an error related to the opposing view. In this mechanism, the fake news creator employs the 'ad hominem' fallacy by attacking the credibility of the opponent instead of addressing the actual topic of discussion. This approach is especially effective in making fake news attractive when a trustworthy journalistic source falters or publishes incomplete data only to later revise an article with updated/newly verified information. As an argumentative strategy, a fake news piece might amplify or distort the significance of a 'mainstream' source publishing inaccurate information - setting up the mainstream as a strawman. This can then be generalised as a sign that all associated mainstream sources are untrustworthy or corrupt, implying that only 'alternative' (i.e., fake) sources can deliver credible information (Zompetti 2019). If readers can be convinced that traditional outlets cannot be trusted, fake news sources can appear more credible or attractive by offering a 'different perspective'. This style of argumentation is particularly effective at undermining the credibility of scientists, experts, and public institutions.

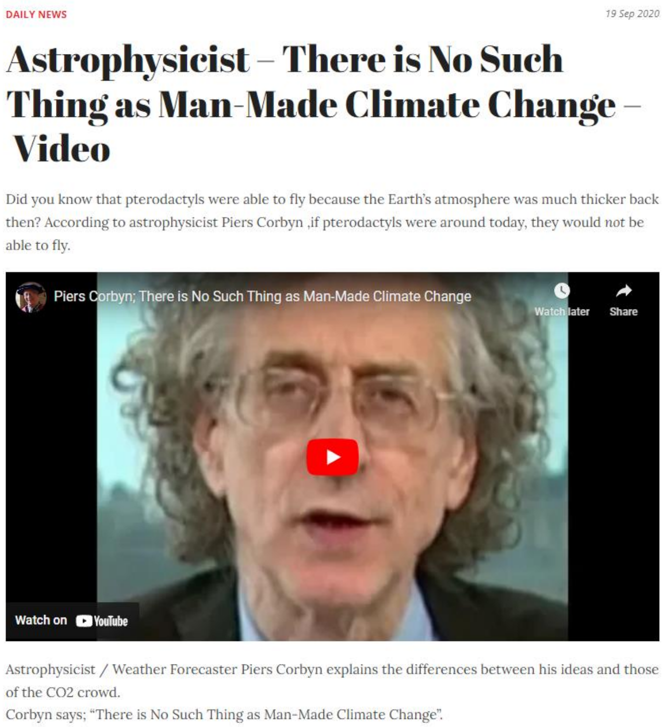
In addition, 'straw manning' can be even more effective if a fake news piece quotes (or misrepresents) their own 'expert' who disagrees with mainstream consensus, claiming to 'prove' opponents wrong by 'cherry-picking' (Musi and Reed 2015) or arbitrarily choosing rare instances where traditional media commits an error. In this way the straw man can be found and set up retroactively using the 'Texas sharp shooter' approach. Generally used in the context of uncovering fallacies, the Texas sharpshooter phenomenon gets its amusing name from folklore about an inept marksman who fires a random pattern of bullets at the side of a barn, and then draws a target around bullet holes, pointing proudly at his success. Similarly in (fake) news, and in science (Nuzzo 2015), authors may pick self-fulfilling data to match their biases and lend credibility to their arguments.

In a classic example of misrepresentation and feigning authority, the example here includes an opinion piece by Piers Corbyn being passed off for "fact" and "science." As *AFP Factcheck*, a fact checking website explains: Piers Corbyn is the brother of UK politician and former Labour Party leader Jeremy Corbyn. The former has spread false claims about climate change for decades. He obtained a degree in physics from Imperial College London in 1968, as well as a postgraduate degree in astrophysics from Queen Mary College in 1981. However, he is not a climate scientist and has very little related scientific research or peer-reviewed papers to his name."[41] So, even though Piers Corbyn is, in fact, a trained scientist, his field does not overlap with that of climatology and his perceived authority or training cannot compensate for that.

Besides, context is important - Corbyn presents climate change as "propaganda," on his weather forecast platform WeatherAction. In a March 2020 tweet, Corbyn falsely claimed the health crisis was a "simulation" by "mega-rich control freaks" Bill Gates and George Soros[42]. In February 2021, he was arrested for comparing vaccination to the Holocaust, according to *The Guardian*[43].

Equivalency and Emphasis Framing

Consistent with the above mechanisms, two other methods used to misrepresent and argue for fake news are equivalency framing and emphasis framing where:



DAILY NEWS 19 Sep 2020

Astrophysicist – There is No Such Thing as Man-Made Climate Change – Video

Did you know that pterodactyls were able to fly because the Earth's atmosphere was much thicker back then? According to astrophysicist Piers Corbyn ,if pterodactyls were around today, they would *not* be able to fly.

Piers Corbyn; There is No Such Thing as Man-Made Climate Change
Watch later
Share
Watch on YouTube

Astrophysicist / Weather Forecaster Piers Corbyn explains the differences between his ideas and those of the CO2 crowd.
Corbyn says; "There is No Such Thing as Man-Made Climate Change".

- ➔ Equivalency framing occurs when the communicator uses an alternate word or phrase to describe an event or issue; while the meaning and the indicated outcome of each term are logically identical, using one term instead of the other results in different preferences among message recipients (Chong and Druckman 2007; Druckman 2001). In simpler words, the same news is adjusted semantically in order to make the argument more preferable or agreeable for different sets of viewers.
- ➔ Emphasis framing, also called value and issue framing, involves the communicator using certain words and concepts when making a statement with the purpose of emphasising

[41] "British Meteorologist Falsely Blames Climate Change on Sun, Moon." Fact Check, 28 Sep. 2022, factcheck.afp.com/doc.afp.com.32JY77E-1.

[42] Archived at: https://web.archive.org/web/20200316014914/https://twitter.com/Piers_Corbyn/status/1239367865784033280

[43] Quinn, Ben. "Piers Corbyn Arrested Over Leaflets Comparing Vaccine Programme to Auschwitz." *The Guardian*, 5 Feb. 2021, www.theguardian.com/uk-news/2021/feb/04/piers-corbyn-arrested-over-leaflets-comparing-covid-vaccine-programme-to-auschwitz.

specific considerations (Druckman 2001; Druckman 2004). As in, framing issues through added emphasis on the values and beliefs of intended audiences, making the news more attractive, and harder to ignore.

Gish-galloping

As an alternative argumentative strategy, a fake news piece may not seek to directly 'convince' or 'prove' a falsehood to readers. Instead, fake news might seek to confuse or disorient readers, bombarding them with contradictory or irrelevant information. Readers may encounter a rhetorical strategy known as "gish-galloping", where a writer or speaker "careens through topics, rattling through half-truth after half-truth… [aiming] both to overwhelm opponents' ability to respond and to introduce doubt into the minds of audiences" (Johnson 2017). This kind of rhetoric can be intimidating for readers and viewers establishing an air of authority for purveyors of misinformation. Gish-galloping may be especially successful in disorienting consumers who feel that they do not have enough background or experience with the topic to challenge what they are seeing.

Fact Signalling

A related tactic involves what Hong & Hermann (2020: 1) call "fact signalling": the "performative invocation of the idea of Fact and Reason". Instead of presenting concrete evidence or using sound reasoning, a fake news piece might condescendingly wield 'facts' as a weapon against perceived opponents. The relevance or truthfulness of these 'facts' is irrelevant; what matters for this strategy is the affective performance of authority and the belittling of opposing viewpoints. Put differently, this is an appeal to emotion masquerading as rationality. Fact signalling appeals to readers by manipulating "what looks like truth, what sounds authentic, [and] what feels reasonable" (Hong & Hermann 2020: 3)

The following quote from Hong & Hermann's paper encompasses risks of fact-signalling concisely: "Scholars are increasingly attentive to the ways in which what was once popularised as a 'fake news' epidemic is not simply a virulent strain of bad information in a fundamentally rational online ecosystem, but rather a broader crisis and transformation of what counts as truthful, trustworthy and authentic (e.g. Boler & Davis, 2018; also see Banet-Weiser, 2012)."

Impression of Expertise

In order to have the desired effect, there is one more piece of the puzzle which must fit, however. This factor involves the ways in which a fake news distributor can assert their credibility or give an impression of expertise. Zyl and colleagues summarise, for example, the 'Checklist for Information Credibility Assessment' put forth by proponents of digital literacy. In their summary, they list parameters such as accuracy, authority, objectivity, currency, and coverage or scope where:

➔ **Accuracy** indicates the degree to which the news content is free from errors - this may include both superficial errors such as spelling or punctuation, as well as errors within the message itself. In addition, accuracy refers to whether the information can be verified elsewhere. In fact, not only is it an indication of the reliability of the information at hand, but also by extension, the reliability of the website or news source itself.

➔ **Objectivity** is an exercise in deciding whether the content being presented is opinion or fact, and whether there is commercial interest, indicated for example by a sponsored link.

➔ Lastly, **coverage** or scope refers to the depth and comprehensiveness of the information presented. One tries to decide if the coverage of the subject at hand is rather superficial, or the author demonstrates an adept understanding of the topic.

According to the proponents of this checklist, therefore, not only does the content matter but also the way in which it is presented (error-free, grammatically correct) and what kind of an impression does it give of the author (do they seem proficient or knowledgeable in the field).

Interestingly, however, in a series of studies conducted by Metzger and her colleagues (2007), it was found that even when supplied with a checklist, users rarely used it as intended (Zyl et al, 2020, 27). Currency, comprehensiveness, and objectivity were only occasionally verified, while checking an author's credentials was the least preferred method of verification by users. This correlates with findings by Eysenbach and Köhler (2002) who indicate that the users in their study did not search for the sources behind the presented website, nor were they interested in learning more about how the presented information was compiled. This lack of thoroughness is ascribed to the users' lack of willingness to expend cognitive effort (Eysenbach and Köhler 2002 29).

These studies are compelling especially in the context of the 'post truth' era where we are constantly bombarded with new 'news'. The unprecedented spread of misinformation begs the question - 'How often and how reliably can we perform an information credibility check in an atmosphere saturated with so-called news?' Following through a mental checklist, however rudimentary, would presumably get tiring if performed with such frequency. In order to minimise cognitive and decision-making effort, people often skip through the arduous task of executing a credibility assessment as shown by the studies mentioned above.

This apparent attempt by users to minimise mental effort has given rise to other studies on how users apply cognitive heuristics as well as other 'short-cut' means to assess news more quickly and with less effort. Sparring research on authority and expertise in the field of fake news has led to the development of a number of descriptive models and theories on how users assess credibility *in practice*.

Discursive Mechanisms

As described in the previous sections, fake news can gain credibility and appeal to readers through narrative mechanisms (involving storytelling and identity) and argumentative mechanisms (involving claims to authority, logic, and rationality). Yet there are still different - and sometimes more subtle - ways in which fake news can attract readers. Discursive mechanisms involve the 'big picture' aspects which affect consumption of fake news. These mechanisms relate to the social and technological contexts in which fake news gets noticed and thrives. We can analyse these mechanisms by examining the form, function, and distribution method of a given fake news piece.

As mentioned in the previous section, the unprecedented increase in the volume of 'news' we encounter today makes it nearly impossible to verify every piece of information we come across. To combat this, people rely on wider mechanisms or *discourses* that form the background for news consumption alongside mental short-cuts also known as cognitive heuristics. Put simply, we rely on shared knowledge and group judgement when judging news which may be false, as in, *'What do others think about this?'*

Reputation, Endorsement and Repeated Exposure

The foundation for "use of cognitive heuristics" approach was established in the 1950s by economist and cognitive psychologist Herbert A. Simon. In refuting the concept of 'rational choice theory,' Simon claimed that although people use reasoning to perform a cost-to-benefit analysis, they can never really determine the true costs or benefits of each action because knowing all the costs and outcomes is a human impossibility; something he called 'bounded rationality'. If we consider bounded rationality as a fundamental feature of cognition, as a consequence, problem solving cannot be exhaustive: i.e., we cannot explore all the possibilities which confront us, and search must be constrained in ways that facilitate search efficiency even at the expense of search effectiveness (Richardson 2017). In simple words, the use of mental shortcuts or heuristics is a method that we apply when faced with a problem, such as deciding whether or not a certain piece of news is fake or real.

Since we physically cannot perform a detailed investigation of every post we see online, we 'satisfice' - or rely on mental shortcuts or 'rules of thumb' until the acceptability threshold is met. Zyl and others (2020) have applied the theory of cognitive heuristics to fake news consumption. According to them, during the examination of an online news article, we go through metrics such as: reputation, endorsement, consistency, expectancy violation and persuasive intent.

→ The **reputation** heuristic may be exercised in many different circumstances. Heuristic may represent literally, the reputation of the source as a news reporter. Or it may call upon tropes of brand loyalty, indicating that the source is a website or brand which the user recognizes or is familiar with. For example, one of the first markers which makes news attractive to readers is its reliability, and one of the best ways to 'guess' whether a piece of news is reliable is that it was featured on a reputed news outlet and/or an outlet that the reader regards highly.

→ The **endorsement** heuristic applies the logic of *'if others believe it, it must be true'*. This may include both groups of people - people whom the user is familiar with such as friends and family, but also other people on the internet who the user doesn't personally know but have given a service or source a review or comment sharing their own experience with the source. E.g., when a user sees news from an unfamiliar region or country, they may look to what others have said - if they can find local support for the source of news - they are likely to find the news valuable and believable even if it contradicts common sense.

→ The **consistency** heuristic indicates that if similar information about an element appears on other sources or websites, it is credible and therefore attractive. For example, if a person hears seemingly unbelievable news from a friend, they may not think much about it. However, if this news later reaches them through another, unrelated source - it may pique the reader's interest, they may find it more attractive than someone encountering this information for the first time.

Fogg's (2002) web credibility framework is another model through which we can understand what makes fake news attractive. It is built on three categories - operator, content and design where:

➔ **Operator** refers to the source of the website - the person who runs and maintains the website. According to Fogg, a user makes a credibility judgement based on the person or organisation operating the website.

An example of this would be when a user is faced with news regarding the discovery of a new planet, the user may not be particularly interested in astronomy but may get interested when they see that the source of this information is NASA. In essence, the user may find the news more attractive due to the perceived trustworthiness of the source.

➔ **Content** refers to the content and functionality of the website where the news is found. Of importance is the currency, accuracy and relevance of the content and the endorsements from external organisations that are deemed respectable.

For example, if a reader comes across a news article on Twitter which was re-tweeted by a reputable news outlet such as the BBC, the original article becomes more attractive, even if the original article was not actually written by the BBC. In this example, the fact that the original writer of the article is unknown or unpopular is compensated for (and made attractive) by the reputation of Twitter as a household name for 'serious' or 'academic' social media rife with debate; as well as by the endorsement of a big-name journalistic channel which is BBC.

In a similar vein, we would like to point out that there is something to be said also about the **channel** through which (dis)information is received. In many cultures, fake news can obtain a layer of legitimacy if received through a trustworthy channel - for example through a family member or someone more educated.

Design and Visual Markers

Penultimately, borrowing again from Fogg's (2002) framework is another superficial but vastly interesting feature - design. Fogg describes design as the structure and layout of the website. Through mutual consensus and with online spaces (particularly social media) increasingly becoming an integral part of our daily lives, we have identified a certain 'look' for what news should look like. Fogg breaks down this element of design into four aspects namely:

- Information design as in how the information is structured on the website, does it make sense, is it logical, does it follow a chronological pattern: etc.
- Technical design, the functioning of the website/source on a technical level, e.g., whether it has a search function; do the hyperlinks work as intended; and so on.
- Aesthetic design speaks to the visual presentation of the source including the looks, feel and professionality of the design. When a website presents news in an appropriately

‘intelligent’ style, using sombre or ‘academic’ colours, the reader is more likely to see the source as smart or attractive.

- Interaction design speaks to the ease of navigation, and interaction with the source as well as the user interface - Is it obvious where the reader must click to move on to the next page? Are all the photos readily visible? Is the post interactive in the sense of including motion or visual aids such as graphs to make the article easy to follow?

Click baiting

The most familiar of these discursive mechanisms, in terms of visual design, can be seen in the use of ‘clickbait’ in fake news pieces. Clickbait, essentially an advertising tactic, describes a “[headline] whose main purpose is to attract the attention of readers and encourage them to click on a link to a particular webpage” (Zhou et al 2021: 2). Effective clickbait titles are outrageous, challenging, combative, even amusing, and tempt individuals scrolling past to click and see what the media has to say. They might offer a dramatic question (e.g., “Have you seen what the Prime Minister has to say about THIS EVENT”?) or only share part of the information - causing a ‘cliff-hanger effect’ (e.g., “Queen Elizabeth Says: “Muslim Refugees Are Dividing Nationality, I Fully Agree With Donald Trump We Should...")[44].

Queen Elizabeth Says: “Muslim Refugees Are Dividing Nationality, I Fully Agree With Donald Trump We Should..."

Cable and Mottershead (2018) have studied the cliff-hanger effect specifically in the context of sports news. They explain: “If a headline features a cliff-hanger, for instance, then we will be inclined to click because we want to find out the answers. It is this feeling of deprivation which provokes the reader into making these decisions.” They offer the following example from *The Guardian*[45]:

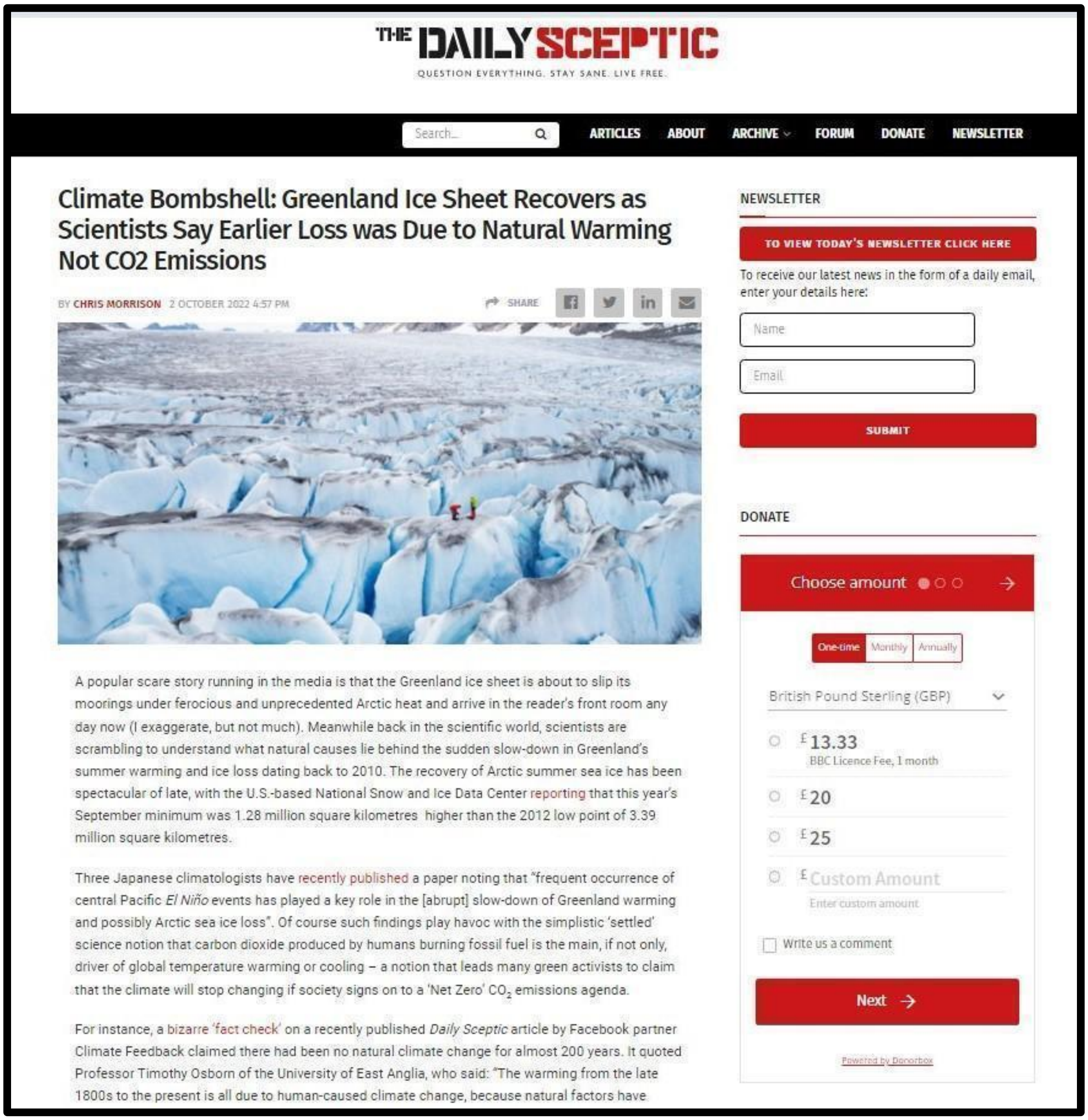
Manchester United: five things we have learned from the US tour

This example is typical of a sports article. It claims to have uncovered new information, but the headline construction is careful not to give any of this new material away. The title is short and tantalising, a cliff-hanger and points towards a knowledge gap as it gives no answers. These traits of ‘clickbait’ make news more attractive as readers look to bridge the distance between their new-found curiosity and promised knowledge.

[44] Archived at: https://archive.ph/3tBtp

[45] https://www.theguardian.com/football/2014/aug/06/manchester-united-five-things-learned-usa-tour

3.3.1 Case study (1) – Climate denial



Climate Bombshell: Greenland Ice Sheet Recovers as Scientists Say Earlier Loss was Due to Natural Warming Not CO2 Emissions

BY CHRIS MORRISON 2 OCTOBER 2022 4:57 PM

A popular scare story running in the media is that the Greenland ice sheet is about to slip its moorings under ferocious and unprecedented Arctic heat and arrive in the reader's front room any day now (I exaggerate, but not much). Meanwhile back in the scientific world, scientists are scrambling to understand what natural causes lie behind the sudden slow-down in Greenland's summer warming and ice loss dating back to 2010. The recovery of Arctic summer sea ice has been spectacular of late, with the U.S.-based National Snow and Ice Data Center reporting that this year's September minimum was 1.28 million square kilometres higher than the 2012 low point of 3.39 million square kilometres.

Three Japanese climatologists have recently published a paper noting that "frequent occurrence of central Pacific *El Niño* events has played a key role in the [abrupt] slow-down of Greenland warming and possibly Arctic sea ice loss". Of course such findings play havoc with the simplistic 'settled' science notion that carbon dioxide produced by humans burning fossil fuel is the main, if not only, driver of global temperature warming or cooling – a notion that leads many green activists to claim that the climate will stop changing if society signs on to a 'Net Zero' CO_2 emissions agenda.

For instance, a bizarre 'fact check' on a recently published *Daily Sceptic* article by Facebook partner Climate Feedback claimed there had been no natural climate change for almost 200 years. It quoted Professor Timothy Osborn of the University of East Anglia, who said: "The warming from the late 1800s to the present is all due to human-caused climate change, because natural factors have

To receive our latest news in the form of a daily email, enter your details here:

In this section, we demonstrate features of fake news by exploring the following example[46]:

In October 2022, a link to this article published by *The Daily Sceptic* (TDS) was posted on Facebook and widely shared on social media platforms including Twitter. The post attempts to refute the position that the unprecedented rise in Arctic Heat during the last few decades has been

[46] Full article is available at: https://perma.cc/PX6H-ZDKG

caused by anthropomorphic climate change. Instead, the article falsely contends that "natural warming" can fully explain the loss of ice in the Arctic.

This article provides a compelling showcase for several types of narrative, argumentative, and discursive mechanisms commonly used in fake news:

The Daily Sceptic's article is **prominent** and visually attractive, designed to catch the attention of social media users scrolling past. The title is a good example of **'clickbait'**: *Climate Bombshell: Greenland Ice Sheet Recovers as Scientists Say Earlier Loss was Due to Natural Warming Not CO2 Emissions*. Sensational cues such as "Bombshell", and tropes of credibility "Scientists Say…" make casual users scrolling past curious, calling to a knowledge gap as pointed out by Cable & Mottershead (2020).

Once the quippy title has brought the reader to TDS website, there are many practical and design elements which come into play. These cues are subtle but encourage the reader to trust what they are seeing. The colour scheme and layout used by the website, or the **information and aesthetic design**, although eye-catching (red), is simple, streamlined, and similar to many legitimate news outlets. In addition, the by-line includes a name likely to be familiar to English language readers, along with a date and a time stamp, providing **currency** and **operator information** as added layers of legitimacy. In both - the post shared on social media platforms and in the website version - a large portion of the screen space (especially on mobile phones) is occupied by the large, demanding photo of an ice sheet - attractive 'evidence' of the recovered ice sheet in Greenland. Once the reader begins to read, they experience **ease of comprehension** - an essential requirement for narrative engagement. The **language** used in the article is itself rather informal, simplified and easy to understand - despite addressing a complex issue as Climate change - something that readers may appreciate and be able to **focus attention** on.

In terms of building the narrative, TDS's article begins with an **emotionally charged** indictment of "the media", suggesting that most reporting on climate change is baseless:

"A popular scare story running in the media is that the Greenland ice sheet is about to slip its moorings under ferocious and unprecedented Arctic heat and arrive in the reader's front room any day now (I exaggerate, but not much)."

The text then describes the "scientific world" as confused and "scrambling" to understand a recent slow-down in Greenland's ice loss. The slow-down in ice loss is real (relating to the *El Niño* phenomenon), but its significance and link with climate change are greatly misrepresented. The authors describe a world where **ill-intentioned 'mainstream'** scientists collude with activists and governments to enact the "emissions agenda". This kind of broad, conspiratorial framing of scientists, governments, and 'the media' is common with fake news. A story is presented here which suggests that institutions cannot be trusted, that the government (along with 'others') want to change society in ways that might harm *you*, the reader. Why tolerate emissions regulations and petrol taxes for an 'agenda' that might make life more expensive for you? Why believe the 'popular' take on climate change when this writer (apparently) proves inconsistencies in climatology? The narrative framing of this article is emotionally engaging, **exploiting a disillusionment** many readers may feel towards governments and experts; specifically in the context of inflation and higher energy costs caused by the aftermath of COVID-19 pandemic and the war in Ukraine.

Even if the reader is not entirely convinced by this narrative framing, they may come away **confused and unsure** what to believe. This article treats academic discourse and anecdotal 'evidence' with the same legitimacy, following up out-of-context quotes from climatologists with vague 'refutations' from other Daily Sceptic posts. The implication is that expert sources or scientific journal articles are no more trustworthy than opinion pieces, social media posts, or personal blog posts. Fake news articles use these **false equivalencies** to create confusion and doubt in readers.

This article presents a clear example of several logical fallacies and appeals that are common in fake news. We can begin with the tagline under the website: *"Question everything. Stay sane. Live Free."* - an attempt to give the **impression of neutrality, apoliticality and stoic skepticism**. When readers read this kind of a tagline, they may see this website as trustworthy, put together by authors interested in finding out the 'real truth' about climate change.

This is a sophisticated example of fake news, as the author *does*, in fact, cite two peer-reviewed sources from Nature (conducted by Japanese climatologists[47]) and Quaternary Science Reviews. At first glance, then, this seems to be a post supported by scientific research. However, the author selectively quotes or **cherrypicks** and misinterprets the content of the scientific articles, employing the **Texas sharpshooter fallacy**. If we follow the article link to the Japanese study, for example, we immediately see a finding that contradicts with The Daily Sceptic's article - "*Both natural variability and anthropogenic forcing contribute to recent Greenland warming by reducing cloud cover*".

It is interesting that The Daily Sceptic's website provides a link to the original study at all. If a reader does not click through to read the original source (or if they are inexperienced in reading scientific articles), it may be easy for them to trust The Daily Sceptic's misrepresentation. An average reader approaching their website, especially through a social media channel, would be very unlikely to click through several pages to get to the study due to the concepts of heuristics and satisficing. As discussed above, a majority of online news consumers are likely to expend as little cognitive effort as possible. In this case, simply seeing that a link to the original article is provided might be enough to **'satisfice'** and **meet the minimum acceptability** threshold. The author from TDS does not cite and link these articles for their scientific content, but instead 'name-drops' them as an appeal to authority, lending the post credibility. In addition to these citations, the author also includes a graph later on in the body of the article as well as several mentions of different Professors and Doctorate holders at prestigious and well-known universities (MIT, University of East Anglia, Manchester Met University and Aarhus) to increase the **endorsement value** of the article.

In discussing climate change, the author also creates a **false dilemma** between anthropomorphic climate change and short-term phenomena like *El Niño*. The author presents these as opposing or competing events, when in fact they have little relation. Climate change does

[47] Link to original Japanese study: https://www.nature.com/articles/s43247-021-00329-x

not imply warming temperatures everywhere all the time, and a year of slowing ice loss does not 'disprove' climate change as a trend. The author uses this false dilemma, **straw manning** or **oversimplifying** the situation, to jump to the erroneous conclusion that corrupt scientists are fighting to "*preserve the fiction*" that human activity is responsible for climate change.

At this point, we have an appealing narrative framing (us vs them, distrust of experts and the 'mainstream', etc) and an argument that relies on logical fallacies. Along with, design elements and discursive support from 'viral' sharers on social media, this article is successful in attracting viewers to land and stay on their web page when browsing from social media. Whether the article actually convinces readers of its narrative or argument is not necessarily measurable, but it is easy to imagine how the article could, at least, successfully sow seeds of doubt in the mind of an unsuspecting reader.

In conclusion, fake news is not a new phenomenon. Yet, the current proliferation of fake news presents a unique challenge, something distinct from 'typical' misinformation in its complexity, distribution, and decentralization. The unprecedented increase in availability of technological devices, internet connections, and online sources of information means that any person who is in possession of a device with an internet connection can potentially become a consumer or distributor of fake news.

This phenomenon makes the current era of fake news particularly challenging. We are faced with the dilemmatic convergence of inclusivity in publishing and increased difficulty in judging credibility of online information. The internet (and social media in particular) has 'democratised' media creation, allowing users to quickly and easily share stories, articles, photos, and videos with others. This has brought clear benefits, as social media can empower individuals to express themselves, organise, and access information in ways that would not have been possible otherwise. Creating and sharing content globally has never been easier. Yet there are also risks and drawbacks with this changing communication landscape. It can be difficult to verify sources of information, and the structure of social media websites incentivises sensationalism and emotional engagement.

Despite extraneous efforts to establish gatekeeping or verifying mechanisms, the fleeting nature of digital information simply does not allow for policing of shared information. And while this lack of policing is an excellent development from the perspective of including public and traditionally marginalised populations - the layperson, the citizen journalist - there is also a wide scope for manipulation of this freedom. As Zyl and colleagues point out, digital content is easy to publish anonymously, and easily plagiarised and altered.

References:

1. Abbott (2002) as quoted in Tamul, Daniel J. and Jessica Hotter. (2019) "Exploring Mechanisms of Narrative Persuasion in a News Context: The Role of Narrative Structure, Perceived Similarity, Stigma, and Affect in Changing Attitudes." Collabra: Psychology
2. Banet-Weiser, S. (2012). Authentic™. In *Authentic™*. New York University Press.
3. Baptista, J. P., & Gradim, A. (2020). Understanding fake news consumption: A review. *Social Sciences*, *9*(10), 185.
4. BARBER, J. F. (2020). Fake News or Engaging Storytelling?. *Radio's Second Century: Past, Present, and Future Perspectives*, 96.

5. Benjamin, D., Por, H. H., & Budescu, D. (2017). Climate change versus global warming: who is susceptible to the framing of climate change?. *Environment and Behavior*, *49*(7), 745-770.
6. Boler, M., & Davis, E. (2018). The affective politics of the “post-truth” era: Feeling rules and networked subjectivity. *Emotion, Space and Society*, *27*, 75-85.
7. Brites, M. J., Amaral, I., & Catarino, F. (2019). The era of fake news: digital storytelling as a promotion of critical reading. In *INTED2019 Proceedings* (pp. 1915-1920). IATED.
8. Turner, V. W., & Bruner, E. M. (1986). The anthropology of experience.
9. Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals’ belief in fake news: A scoping review determinants of belief in fake news. *PLoS One*, *16*(6), e0253717.
10. Burkhardt, J. M. (2017). History of fake news. *Library Technology Reports*, *53*(8), 5-9.
11. Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media psychology*, *12*(4), 321-347.
12. Cable, J., & Mottershead, G. (2018). 'Can I click it? Yes you can': Football journalism, Twitter, and clickbait. *Ethical Space*, *15*(1/2).
13. Druckman, J. N. (2001). The implications of framing effects for citizen competence. *Political behavior*, *23*, 225-256.
14. Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *Bmj*, *324*(7337), 573-577.
15. Fogg, B. J. (2003, April). Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems* (pp. 722-723).
16. Hong, S. H. (2020). “Fuck Your Feelings”: The Affective Weaponization of Facts and Reason. In *Affective Politics of Digital Media* (pp. 86-100). Routledge.
17. Horne, B. D., & Adali, S. (2017, May). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh international AAAI conference on web and social media*.
18. Introne, J., Iandoli, L., DeCook, J., Yildirim, I. G., & Elzeini, S. (2017, July). The collaborative construction and evolution of pseudo-knowledge in online conversations. In *Proceedings of the 8th International Conference on Social Media & Society* (pp. 1-10).
19. Johnson, A. (2017). The multiple harms of sea lions. *Harmful Speech Online*, 13.
20. Machete, P., & Turpin, M. (2020). The use of critical thinking to identify fake news: A systematic literature review. In *Responsible Design, Implementation and Use of Information and Communication Technology: 19th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2020, Skukuza, South Africa, April 6–8, 2020, Proceedings, Part II 19* (pp. 235-246). Springer International Publishing.
21. Meneses, J. P. (2018). Sobre a necessidade de conceptualizar o fenómeno das fake news. *Observatorio (obs*)*, (1), 37-53.
22. Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology*, *58*(13), 2078-2091.
23. Mikołajczak, M., & Bilewicz, M. (2015). Foetus or child? Abortion discourse and attributions of humanness. *British Journal of Social Psychology*, *54*(3), 500-518.
24. Musi, E., & Reed, C. (2022). From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*, *33*(3), 349-370.
25. Nuzzo, R. (2015). Fooling ourselves. *Nature*, *526*(7572), 182.
26. Pengnate, S. F., Chen, J., & Young, A. (2021). Effects of clickbait headlines on user responses: An empirical investigation. *Journal of International Technology and Information Management*, *30*(3), 1-18.
27. Polletta, F., & Callahan, J. (2019). Deep stories, nostalgia narratives, and fake news: Storytelling in the Trump era. *Politics of meaning/meaning of politics: Cultural sociology of the 2016 US presidential election*, 55-73.
28. Quinn, Ben. “Piers Corbyn Arrested Over Leaflets Comparing Vaccine Programme to Auschwitz.” The Guardian, 5 Feb. 2021, www.theguardian.com/uk-news/2021/feb/04/piers-corbyn-arrested-over-leaflets-comparing-covid-vaccine-programme-to-auschwitz.

29. Ravaja, N., Saari, T., Kallinen, K., & Laarni, J. (2006). The role of mood in the processing of media messages from a small screen: Effects on subjective and physiological responses. *Media Psychology*, *8*(3), 239-265.
30. Richardson, R. C. (2017). Heuristics and satisficing. *A companion to cognitive science*, 566-575.
31. Ryan, M. L., Ruppert, J., & Bernet, J. W. (Eds.). (2004). *Narrative across media: The languages of storytelling*. U of Nebraska Press.
32. Schuldt, J. P., Konrath, S. H., & Schwarz, N. (2011). "Global warming" or "climate change"? Whether the planet is warming depends on question wording. *Public opinion quarterly*, *75*(1), 115-124.
33. Stein, N. L. (1982). What's in a story: Interpreting the interpretations of story grammars. *Discourse Processes*, *5*(3-4), 319-335.
34. Tandoc, E. C., & Seet, S. K. (2022). War of the Words: How Individuals Respond to "Fake News,""Misinformation,""Disinformation," and "Online Falsehoods". *Journalism Practice*, 1-17.
35. Watson, C. A. (2018). Information literacy in a fake/false news world: An overview of the characteristics of fake news and its historical development. *International Journal of Legal Information*, *46*(2), 93-96.
36. Pengnate, S. F., Chen, J., & Young, A. (2021). Effects of clickbait headlines on user responses: An empirical investigation. *Journal of International Technology and Information Management*, *30*(3), 1-18.
37. Zompetti, J. P. (2019). The Fallacy of Fake News: Exploring the Commonsensical Argument Appeals of Fake News Rhetoric through a Gramscian Lens. *Journal of Contemporary Rhetoric*, *9*.
38. Van Zyl, A., Turpin, M., & Matthee, M. (2020). How can critical thinking be used to assess the credibility of online information?. In *Responsible Design, Implementation and Use of Information and Communication Technology: 19th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2020, Skukuza, South Africa, April 6–8, 2020, Proceedings, Part II 19* (pp. 199-210). Springer International Publishing.

3.2 Critical Thinking

Ruxandra Buluc, Cătălina Nastasiu

Abstract

The present section investigates critical thinking as one of the most recommended instruments in the fight against disinformation. Numerous European documents and courses point to developing critical thinking as a means of countering the effects disinformation has on the informational environment. The section examines the skills that need to be included in critical thinking curricula, the traits of a critical thinker, the processes and types of analysis involved in critically interacting with any kind of discourse (argumentative, rhetorical, conversational, narrative). The deliverable also puts forth concrete tools for students to develop their critical thinking competences. Employing the latest research in the field, the section focuses on the skills and competences that need to be developed in order for students to become proficient critical thinkers, providing them with clearly understandable checklists and questions, as well as examples and analyses to help them understand how critical thinking works.

Main research questions addressed

- What is critical thinking?
- What abilities make a person and efficient critical thinker?
- What are the types of analysis which comprise critical thinking?
- How can each type of analysis be developed?

European documents are increasingly mentioning critical thinking as a solution to counter the effects of disinformation and to build citizen resilience in the face of disinformation attacks and foreign information and manipulation interference.

The *Digital Education Action Plan* presents critical thinking as a requirement in today's society and combines it with media literacy to develop the "ability to engage positively and competently in the digital environment" (2018, 3) and to "overcome the ever-present threats of fake news, cyber bullying, radicalisation, cybersecurity threats and fraud" (2018, 8). However, the document acknowledges the fact that making these abilities available to the wider public still proves somewhat elusive.

The *Communication on achieving the European Education Area by 2025* names critical thinking as one the key transversal skills that the next generations of students should possess, along with entrepreneurship, creativity and civic engagement through transdisciplinary, learner-centred and challenge-based approaches (2020, 5).

In the context of digital rhetoric, critical content interpretation capabilities are essential. Critical thinking represents a set of tools used in the process of interpreting content to analyze and evaluate ideas, messages, or arguments, including interpreting evidence, placing the message in a larger context, and understanding whether cited data supports the point of view.

The *New Skills Agenda for Europe* also refers to critical thinking as a transversal skill and key competence that needs to be mastered along with digital competences, entrepreneurship, problem solving or learning to learn, and financial literacy (2016, 5). The *New Skills Agenda for Europe* also refers to the broad range of skills which formal education and training should equip graduates with: literacy, numeracy, science and foreign languages so as to foster inclusion, personal development, employability and active citizenship. In the context of current technological and informational changes, critical thinking have become essential for full participation in society. Moreover, new technical and "soft" skills are gaining more and more importance on the labor market.

The first steps in order to provide a curriculum to develop critical thinking are to understand what critical thinking is, what makes a good critical thinker and what skills and types of analysis it relies on.

What is critical thinking?

Walton (1989, 169) explains that critical thinking requires a central theory of reasoned argument criticism. Siegel (1989, 21) points out that critical thinking is “principled thinking” as it combines reasons and principles and it is dependent on, stemming from and promoting rationality, by bringing to light the matters which are relevant for belief formation and action. Hunter (2009, 2) explains that critical thinking is reasonable, reflective thinking that is aimed at deciding what to do and what to believe. Paul and Elder (2020) provide a more extensive definition:

> Critical thinking is the art of analyzing and evaluating thought processes with a view to improving them. Critical thinking is self-directed, self-disciplined, self-monitored, and self-corrective thinking. It requires rigorous standards of excellence and mindful command of their use. It entails effective communication and problem solving abilities, as well as a commitment to overcoming our native egocentrism and socio-centrism. It advances the character and ethical sensitivities of the dedicated person through the explicit cultivation of intellectual virtues.

Their definition examines not only the characteristics of critical thinking as a process, but also its prerequisites (the standards), its outputs (effective communication and problem-solving) and its effects (character molding and evolution). Siegel (1989, 23) also points out that principled critical thinking is based on consistency, impartiality, non-arbitrariness, fairness, which recognises the universal and objective standards on which judgements are based. He (Siegel, 1989, 23-24) details the two types of principles that exist: a) subject-neutral principles, which are general and function across a wide variety of contexts and subjects, such as proper inductive and deductive reasoning, fallacy avoidance; b) subject-specific principles which apply only to specific areas of inquiry or subjects, such as physics, mathematics, linguistics, etc. Both types of principles are necessary prerequisites for effective critical thinking. It is not sufficient to know various facts in a particular subject if one is not able to operate with them logically; just as it is not enough to understand reasoning processes if one does not know the specific knowledge in a field that those processes could be applied to. More often than not, lack of knowledge in one of these areas leads to fallacies, to acceptance of dubious theories (be they fake news, unfounded rumors or conspiracy theories), to a lack of understanding of how the world and society function, which does not allow for a rational and constructive debate in the public sphere (see chapters 2.3 and 3.1 for more information). Subject-specific principles do not form the object of this chapter, as it would be impossible to cover all the different areas. However, we will delve into the subject-neutral principles and skills that help form and inform good reasoning practices across disciplines of study. Critical thinking researchers and educators have identified a series of abilities, skills and sub-skills that need to be developed and practised in order to foster the development of critical thinkers.

a) empathy - the ability to constructively understand the other side's point of view (Walton, 1989, 169)

b) critical detachment - the ability to detect bias, and thereby to avoid being too heavily partisan to attain a balanced perspective in argument.' (Walton, 1989, 169)

To these Facione (1990a, 12–19) adds a further set of skills and sub-skills:

c) interpretation: categorization, decoding significance, clarifying meaning;

d) analysis: examining ideas, identifying arguments, analyzing arguments;

e) evaluation: assessing claims, assessing arguments;

f) inference: querying evidence, conjecturing alternatives, drawing conclusions;

g) explanation: stating results, justifying procedures, presenting arguments;

h) self-regulation: self-examination, self-correction.

To this extended list, we would add that it is not only arguments that serve as the basis for the practice and development of these critical thinking skills; one should also focus on discourse that is not argumentative, on narratives, speech and conversation, as well as on source analysis. All these will be explored in more depth in the following sections.

Therefore, critical thinking can be seen as a process which focuses on the self, as much as on the others in order to identify, monitor, assess, (if necessary) correct the reasons and principles that function in decision-making situations or in communicative contexts more generally. Most of the problems that people face in their everyday lives require thoughtful consideration and cannot be answered with a simple yes or no. These issues may not have clear-cut solutions, may require updated solutions, as society changes quickly and fundamentally. Any and all solutions that may present themselves have both advantages and disadvantages, that necessitate a reasonable, principled, overt analysis, which accounts for various perspectives, uncovers possible omitted elements, unearths the values and interests that may dictate opinions, accounts for the contextual influences and limitations. This is the main task of critical thinkers and this approach to situations, events, opinions, convictions, beliefs, values presupposes that people think independently, can analyse with detachment their own reasoning as well others', and can make informed, calculated and responsible decisions.

What makes a critical thinker?

Critical thinkers exhibit a complex set of traits, attitudes and dispositions that Siegel (1989, 21-22) groups under the more general label "critical attitude" or "critical spirit". First and foremost, they strive to determine, identify, evaluate the reasons, based on principles (see section 1 of this chapter) that support any claim, judgement or action. However, this is not sufficient. The critical thinker must also be willing to endorse the principles, and strive to put them into practice through reasoning, have an inquiring mind and a commitment to objectivity as much as possible, sympathetic but thorough investigative inclinations that are not solely focused on the points of view that the thinker supports, but even more so on the ones they are more inclined to reject, so as to be able to overcome some possible biases (see chapter 2.1).

Facione (1990, 25) undertook an extensive mapping of the critical thinkers' skills and attitudes with respect to life in general and also focused on how they could be applied to specific situations or issues. The table below presents the list of skills, aptitudes and characteristics that make adept critical thinkers.

Life in general	**Specific issues**
● inquisitiveness about a wide range of issues; ● concern to become and remain generally well-informed; ● alertness to opportunities to use critical thinking; ● trust in the processes of reasoned inquiry; ● self-confidence in their own ability to reason; ● open-mindedness to divergent worldviews, ● flexibility in weighing alternatives and opinions, ● understanding different opinions; ● fair-mindedness in assessing reasoning; ● honesty in facing their own biases, prejudices, stereotypes, egocentric or sociocentric tendencies; ● prudence in suspending, making or altering judgments, and ● willingness to reconsider and revise views where honest reflection suggests that change is warranted.	● clarity in stating the question or concern; ● orderliness in working with complexity; ● diligence in seeking relevant information; ● reasonableness in selecting and applying criteria; ● care in focusing attention on the concern at hand; ● persistence though difficulties are encountered, and ● precision to the degree permitted by subject and circumstances.

Table 5 Features of critical thinkers

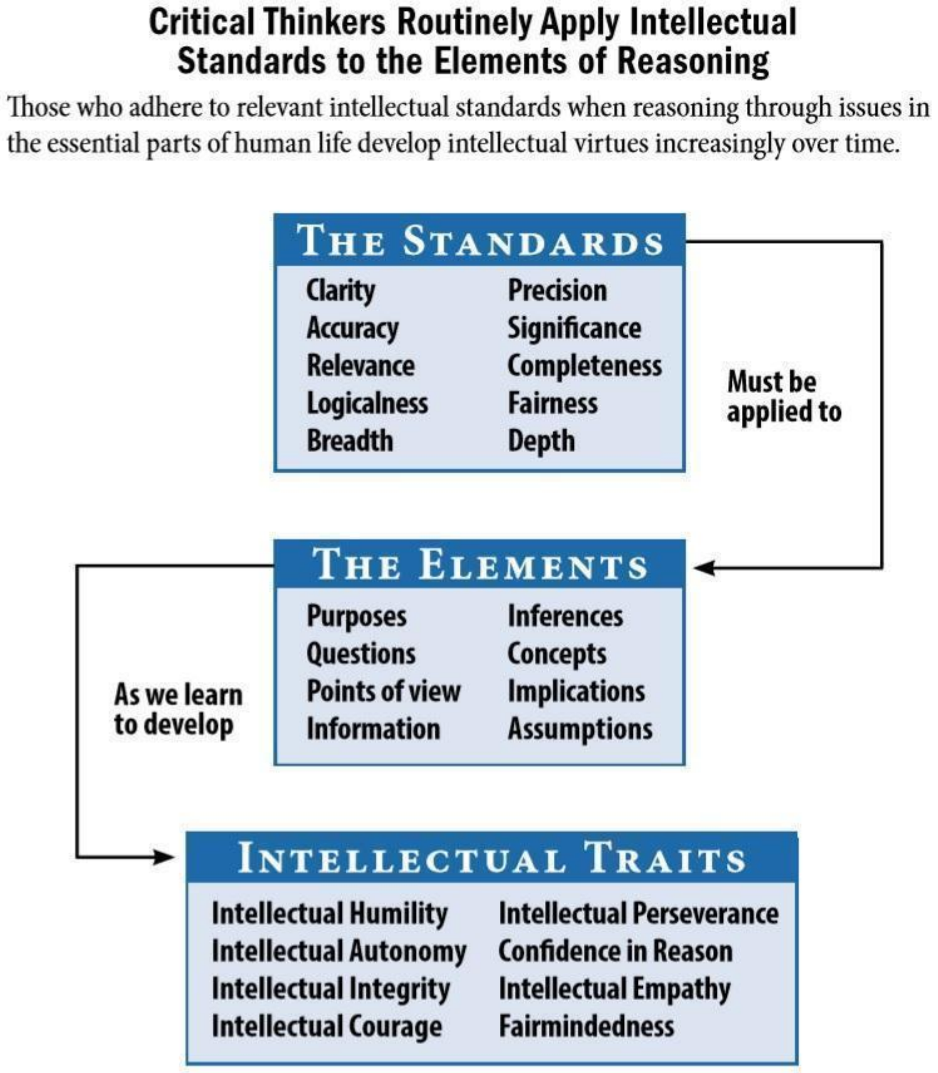
This is a wide array of skills and traits that a person must possess and exercise in order to become an efficient critical thinker. Some of them might be more easily attainable, while others might be more difficult to exercise in every context, due to emotional interferences or charges and also to time constraints that might hinder more in-depth analysis. However, if practised extensively, the skills become more like second nature and manifest themselves when needed, thus turning a person into what Paul and Elder (2020) coined the term "well-cultivated critical thinker".

They concentrated their vast experience in teaching and researching critical thinking and developed a brief toolkit to encompass the skills employed by the well-cultivated critical thinker:

- raises vital clearly formulated questions and issues;
- gathers and assesses relevant information;
- comes to well-reasoned conclusions and solutions and verifies them against principles and standards;
- thinks open-mindedly within alternative systems of thought;
- communicates effectively with others to identify solutions to complex problems; and
- avoids misrepresenting or distorting information in producing arguments (Paul and Elder, 2020, 9).

Hunter (2007) synthesizes the traits that a critical thinker should have under two encompassing categories: reasonableness and reflectivity. The former refers to employing and relying on reason and sound reasoning principles when analyzing and evaluating the arguments and discourses one is presented with. The latter means that the analysis process should be in-depth, exploratory and multi-level, which means it "involves thinking about a problem at several different levels or from several different angles all at once, including thinking about what the right method is for answering or solving the problem (Hunter, 2007, 5). Levitin states that employing reasonableness and reflectivity fosters the development and maintenance of intellectual humility: "Critical thinking trains us to take a step back, to evaluate facts and form evidence-based conclusions. (...) The most important component of the best critical thinking that is lacking in our society is humility. It is a simple yet profound notion: If we realize we don't know everything, we can learn. If we think we know everything, learning is impossible" (2017, xiv). Being able to acknowledge that we may not know everything, that some things are beyond our understanding is the defining quality of a rational thinker, who is able not only to assess the limits of others' judgements but also one's own.

Paul & Elder (2004, 5) also explain that human objectivity and all-knowingness is merely an ideal that individuals cannot attain on their own and that the best ways to try to get as close to objectivity as possible is to remain humble, admitting one's subjective point of view and considering diverging, different, various competing sources of information when considering an important judgment or decision.



Critical Thinkers Routinely Apply Intellectual Standards to the Elements of Reasoning

Those who adhere to relevant intellectual standards when reasoning through issues in the essential parts of human life develop intellectual virtues increasingly over time.

Figure 7. Critical Thinking Process. Source *The Miniature Guide to Critical Thinking Concepts and Tools*, Paul and Elder, 2020, 12

What should critical thinking courses focus on?

Most critical thinking courses focus on formal and informal logic, on producing correct and valid deductive and inductive arguments, which might, at times, appear to students to be cut off from the real world and all the issues that they need to consider. It is our contention that argumentation is but a part, albeit vital, of what critical thinking courses should address. Not all the real-life situations in which people need to exercise critical thinking can be reduced to arguments. In fact, in many cases, people have to analyse other types of discourse, such as persuasive discourse, which is not constructed around an argument; or narratives, which present a message that is not argumentative; or participate in conversations or dialogues that do not revolve around an argument; or, increasingly more important at present, analyse the sources of various messages to determine their credibility and trust-worthiness. Therefore, we argue that critical thinking courses should extend their area of applicability beyond argumentation to also include: rhetorical analysis, narrative analysis, conversational analysis and source analysis (which is explored in more depth in a separate section 3.4 Media literacy).

a) **Argument analysis**

Continuing the discussion of argumentative fallacies in section 3.1, we focus on how critical thinking can aid people to detect, analyse and uncover such fallacies in disinformation and thus counter their manipulative effects.

Critical thinking focuses on explicitly analysing arguments employing the standards and principles that make it a fair, transparent, impartial skill. The way to do this is to uncover the reasons that an argument is built on and to examine their validity. Hunter (2009: 4) explains that there are three important categories of reasons: epistemic, pragmatic and emotional.

Epistemic reasons refer to facts, to the real world, to the system of knowledge and data that society is based upon, that is true regardless of whether people believe them or not. For example, the law of gravity dictates that objects fall to the ground due to gravity irrespective of whether people know about this law, believe it or agree with it. Epistemic reasons are independent of other types of reasons and they confer truth value to various assertions given the scientific evidence and previously proven theories that they encompass. They remain true and real irrespective of how well they are integrated or made use of in an argument. Examples of such reasons are appeals to authority (experts, laws, scientific tenets, etc.) and results of scientific research that have been verified and validated in the scientific community.

Pragmatic reasons refer to the beliefs, information, data, evidence that people believe will assist them in attaining their particular goals with more ease or in a timelier manner. These reasons reflect people’s longer or shorter term expectations with respect to how a system should work, how losses can be prevented or minimised, how gains could be enhanced. In other words, they are interest-based and convention-enforced. They regulate how a community or social system functions in order to meet the expectations or satisfy the needs of the individuals that are part of it. Examples of pragmatic reasons could be examples of particular situations and how events unfold, analogies that help the audience understand how something functions by comparing it to another, already familiar and similar item, etc.

Emotional reasons are based on personal experiences, feelings, sentiments, beliefs that are part of and determine an individual’s identity, and which can affect the ways in which that individual perceives the events around them. Emotional reasons are versatile and possibly the most persuasive, as people are emotionally attached to their convictions and are molded by their personal experiences, and do not react entirely rationally when these are called into question. It is difficult to relinquish beliefs that are fundamental to how one perceives and understands the world and their place in it. Moreover, these emotional reasons can also form the basis for communities, be they large or small, in which case they are consistently reinforced by mere inclusion in the respective community. These types of reasons might be the most challenging to untangle, to bring to light and to objectively analyse. They might also be the most resistant to the open-minded approach that critical thinking presupposes because they are so deeply entwined with personal and collective identities.

Moreover, Kahane (1989, 142) explains that students need an understanding of types of reasons, as well as of the ways in which they may be corrupted through sophistic reasoning. To this end, informal logic, employing real-life examples of fallacies, which are then critically analysed and their flaws revealed, better assist students in tackling argumentative fallacies weaponised into fake news and disinformation. Kahane also points out that critical thinking is a how-to type of course, which teaches students not only how to analyse others' arguments but also how to critically analyse their own arguments, assisting them in taking a step back and clearing their own reasoning and strengthening it (also see section 3.1).

Govier (1989, 117) explains that critical thinking is more extensive than mere argumentation. Other types of reasoning processes are involved, some of them may never become public, but they are nonetheless important and need to be examined critically: hypothesising, deliberating, judging, estimating, investigating similarities and differences, explaining, classifying, interpreting, fact presenting, etc. All these thinking processes may be evaluated via critical thinking mechanisms.

Moreover, critical thinking should not and cannot be limited to analysis of argumentative language. It should also include other modes of representation (images, films, visual representations, etc.), as well as other contexts of language use such as conversations, speeches, narratives.

b) **<u>Rhetorical analysis</u>**

Public discourse is a social process through which the speaker creates and transmits meanings and representations of the world, builds the social image that they want to put forth for the audience. None of us have direct access to reality as such, and each of us perceives it differently. Based on what we experience and the representations we construct, we create and transmit our understanding of the world, while at the same time challenging, changing, shaping, molding the society we live in. In order to understand how public discourse shapes the world, the communities and societies we live in, the critical thinker needs to analyse the ways in which rhetoric is manifest in various contexts such as politics, law, science, social science, journalism, history, public relations, strategic communication, marketing, advertising, education, health, etc.

Rhetorical analysis focuses on public discourse in all its forms, examining in more detail how it is constructed, with a special focus on the rhetorical devices that grant it persuasive power and capture the attention of the audiences. Rhetorical analysis calls for a piecemeal integration of various analysis criteria, focusing on broader mechanisms such as the ways in which discourse creates emotional connections to its audiences, either by employing narratives which stir the audience's reactions, or by relying on argumentation, as previously explained. These broader mechanisms stem from the elements of persuasion that Aristotle first put forth in Rhetoric: logos, pathos, ethos, and not to forget kairos.

Logos refers to the rational arguments that are presented in support of an idea.
Pathos refers to the emotions that are stirred in order to gain support for that particular idea.
Ethos refers to the integrity and responsibility of the persuader, as it is known by the audience, and which thus affects the ways in which the audience listens to and accepts the persuasive message.
Kairos, the least explored of the persuasion components put forth by Aristotle, was clarified by the analysis undertaken by Kinneavy (1986, 80) who defines it as “right or opportune time to do something or right measure in doing something,” thus bringing into clearer focus how reliant on the right moment and the right extent of scope or endeavour persuasion is. Moreover, in 2000, Kinneavy & Eskin (2000) employed computer-assisted discourse analysis and identified the two contexts in which Aristotle defined rhetoric: (1) “Its function is not so much to persuade as to find in each case the existing means of persuasion”; (2) “Rhetoric may then be defined as the faculty of discovering the possible means of persuasion in reference to any subject whatever” (Kinneavy & Eskin, 2000, 434). Given all these components and the ways in which they interact in order to construct a persuasive discourse, it is of great valour for critical thinkers to focus their attention on rhetorical analysis of discourse, in order to uncover the ways, means and interests which the orator may employ in order to convince their audience.

Rhetorical devices are the tactical tools employed to convey a meaning persuasively and attractively to a target audience. Their goal is to assist the target audience in visualising and conceptualising the delivered message, by stimulating the audience’s imaginations, stirring emotions, creating memorable images.

The table below presents the most important rhetorical devices as well as relevant examples for each:

	Rhetorical device	Definition	Example
1.	analogy	indicates similarities between the features of objects, situations, events, persons and indicates that what is applicable to one instance, can be transferred to the other	“Stereotypes about racism, religion, gender or anything else, they’re like cancer. If you had a tumour, you wouldn’t quietly hope that it slowly disappears. You would zap it with chemotherapy, and cut it out, an try every experimental treatment until it was gone. This is no different.” Arnold Schwarzenegger 17.08.2017
2.	antithesis	concepts, ideas, events, positions, situations are placed in a clear-cut opposition, so that only one option seems acceptable	“Now is the time to rise from the dark and desolate valley of segregation to the sunlit path of racial justice. Now is the time to lift our nation from the quicksands of racial injustice to the solid rock of brotherhood.” Martin Luther King Jr., I Have a Dream speech

3.	alliteration	represent the excessive repetition of a sound or group of sounds in the same sentence to draw the attention	“We must choose between greatness or gridlock, results or resistance, vision or vengeance, incredible progress or pointless destruction.” President Donald Trump, State of the Union speech, 2019
4.	anaphora	the repetition of the same word or syntagm at the beginning of several, consecutive paragraphs/sentences to emphasise an idea	“I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident, that all men are created equal." I have a dream that one day on the red hills of Georgia, the sons of former slaves and the sons of former slave owners will be able to sit down together at the table of brotherhood. I have a dream that one day even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppression, will be transformed into an oasis of freedom and justice. I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character. I have a dream today!” Martin Luther King Jr., I Have a Dream speech
5.	personification	transfers human characteristics to inanimate objects or phenomena	The financial markets *fought* hard to secure their gains this week.
6.	rhetorical question	questions that an answer to is not expected or is implied in the way they are formulated	“Are we a nation that tolerates the hypocrisy of a system where workers who pick our fruit and make our beds never have a chance to get right with the law? Or are we a nation that gives them a chance to make amends, take responsibility, and give their kids a better future?” President Barack Obama speech on migration, 20.11.2014

7.	hyperbole	deliberate exaggeration for impact	“The only thing we have to fear is fear itself.” Franklin D. Roosevelt Inaugural address, 04.03.1933
8.	repetition	a word or several words are repeated to increase their impact and retention	“We shall go on to the end, we shall fight in France, we shall fight on the seas and oceans, we shall fight with growing confidence and growing strength in the air, we shall defend our Island, whatever the cost may be, we shall fight on the beaches, we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills.” Prime Minister Winston Churchill, 04.06.1940
9.	metaphor	symbolically representing or associating concepts, ideas that would not or have not been associated before to draw attention	“This invasion of others is a raw material, efficiently and ruthlessly mined, packaged and sold at a profit. A marketplace has emerged where public humiliation is a commodity, and shame is an industry.” “Public shaming as a blood sport has to stop, and it's time for an intervention on the internet and in our culture.” “Online, we've got a compassion deficit, an empathy crisis.” Monika Lewinsky, The Price of Shame, TED talk
10.	humour	unexpected associations, plays upon words, plot twists, situation reversals that generate laughter	“The first thing I would like to say is ‘thank you.’ Not only has Harvard given me an extraordinary honour, but the weeks of fear and nausea I have endured at the thought of giving this commencement address have made me lose weight. A win-win situation! Now all I have to do is take deep breaths, squint at the red banners and convince myself that I am at the world’s largest Gryffindor reunion.” J.K. Rowling Harvard commencement speech 05.06.2008
11.	irony	Something is stated directly, by a different meaning, an evaluation,	“The Green New Deal calls for the elimination of all airplanes. This might seem merely ambitious for

		a judgment, an opposing view is transmitted indirectly and needs to be decoded properly by the recipient.	politicians who represent the densely populated northeast. But how is this supposed to work for our fellow citizens who don't live between Washington and Boston? In a future without air travel, how are people supposed to get around the vast expanses of, say, Alaska during the winter? Tauntauns: a beloved species of repto-mammals native to the ice planet of Hoth. While not as efficient as planes or snow-mobiles, these hairy, bipedal space lizards offer their own unique benefits. Not only are tauntauns carbon-neutral, but according to one report "a long time ago" and "far, far away," they may even be fully recyclable for their warmth on especially cold nights." Senator Robert Lee, Remarks on the Green New Deal, 2019.
12.	litote	deliberate understatement for impact	"The planet does not need us to "think globally, and act locally" so much as it needs us to think family, and act personally." Senator Robert Lee, Remarks on the Green New Deal, 2019.
13.	paralepsis	Introducing an idea, concept, event, etc. while at the same time claiming that it will not be discussed, thus allowing the speaker not to assume responsibility for said idea, concept, event etc.	"I find it interesting that it was back in the 1970s that the swine flu broke out then under another Democrat President, Jimmy Carter. I'm not blaming this on President Obama, I just think it's an interesting coincidence." Michele Bachmann, 28 April 2009
14.	euphemism	Replacing a negatively associated phrase with something neutral or positive sounding	"alternative facts" Kellyanne Conway January 22, 2017

Table 6. Rhetorical devices: definitions and examples

Govier (1989, 122) explains that what critical thinking students require is a check-list of "discrediting factors" that they could identify in any type of public discourse, not only in argumentation, and which could indicate that there are reasoning issues underlying that particular discourse. The list could include factors such as:

- use of emotive language, meant to elicit an emotional rather than a rational response;
- explanations for phenomena that are not real (the earth is flat) or that are flawed;
- classifications that are based on flawed criteria;
- exaggerations that are presented as the only possible explanations;
- unsupported claims, in which case no evidence is offered to support that particular claim;
- alternative positions are not examined;
- rhetorical questions are formulated in support of various claims, instead of justifications or evidence;
- false dichotomies;
- no appeal to authority or to research results is employed to support various claims.

To these we could add:

- the use of intentionally ambiguous language that does not actually state anything clearly;
- the use of complicated jargon to deter understanding and make the discourse seem scientifically sound;
- the use of buzzwords that reflect issues which people consider important, relevant and current but which actually do not reveal anything of substance;
- the use of smokescreens by which a key point is hidden behind a plethora of irrelevant words.

All these rhetorical devices are important for critical thinkers to understand and evaluate, not because they could mask disinformation in all cases, but because they might be employed in order to detract their attention, to persuade them with respect to an issue which is not sufficiently supported by actual evidence, to elicit emotional rather than rational responses, and, generally speaking, to create the impression that a lot has been said on a certain issue and therefore no further discussion is warranted.

c) <u>Conversation analysis</u>

Walton (1989, 174-175) explains that there are several types of dialogue or conversation that are frequently the subject of critical investigations:

1) **persuasion dialogue** in which case each participant tries to persuade the other that their position, assertion, etc. is the best and it should be accepted, based on the premises which are presented. This could devolve from constructive persuasive dialogue into a dispute into which the two sides negate each other and one has to be completely abandoned in order for the other one to be accepted. Examples: parliamentary debates; criminal trials.
2) **inquiry** meant to obtain further information and knowledge on a certain issue, thus gaining the necessary evidence to validate a conclusion. The inquiry is based on obtaining factual premises that demonstrate the conclusion.

3) **negotiation** is based on promoting one’s interests and conceding the least by obtaining the best deal possible. Negotiation does not require convictions, values, ideas, truths, but rather trade-offs and exchanges in order to obtain the best outcome for oneself.

In all of these cases, critical thinking provides a useful toolbox for the analysis of the claims, assertions, persuasive instances, trade-offs, information, facts presented so as to determine the best course of action or whether the dialogues proceed correctly, transparently, openly, responsibly. However, in many instances, analysing conversations and dialogues could present certain challenges, as they usually take place in real time and might overburden the critical thinker. Govier (1989 123) explains that it is sometimes more facile to analyse written discourse from a critical perspective than it is to apply the same methods to conversations and/or spoken language. Several reasons are presented for this difficulty:

- spoken discourse needs to be analysed faster, the elements that make it up, arguments, rhetorical devices, instances of persuasion, emotional appeals etc. must be identified as the speech or the conversation progresses, which requires more adept, attentive critical thinkers;
- social and personal factors, such as social status and power relations, play a role in this type of analysis as they could affect the critical thinker’s acuity;
- real contexts are fast-paced, images may disappear before they are properly interpreted, the conversation may move on to other points, and thus make effective analysis challenging.

There are also some advantages to direct conversational analysis:

- in face-to-face conversations, speakers could be interrogated further with respect to the claims they are making and the arguments they are presenting, to their assumption and unstated beliefs;
- body language and facial expressions could also contribute to evaluating a person’s credibility and legitimacy in making the claims, their honesty, openness and reliability.

However, public debates require engaged, critical thinkers who can spot reasoning flaws and counter them appropriately. Therefore, courses in critical thinking should also explore this type of interactions so as to better assist students in developing the skillset that they require. Bailin and Battersby (2021, 32) suggest engaging the students in an inquiry-based approach to teaching critical thinking, as they explain that arguments do not exist in isolation and they are often part of debates, which come with their own set of rules and expectations. Reasoned judgment and evaluation in such cases requires that the critical thinker is aware not only of the exchanged arguments, statements, claims, etc. but also of the larger social context in which the debate occurs so that they have all the necessary information to evaluate what is being included, and equally importantly, what is left out of the debate, to weigh the merits of what is stated, based on a relevant set of criteria, while keeping one’s mind open to both sides of the debate. They suggest a set of questions that could direct such an inquiry into the solidity of a debate:

- What is the issue approached?
- What claims or judgements about the issue are made?
- What are the reasons and arguments that each side presents?
- What is the relevance of the reasons and arguments to the issue?

- What is the context in which the issue is discussed?
- How strong are the arguments, explanations, hypotheses, data presented by each side?
- What conclusion could the critical thinker draw on their own from what was presented? (Battersby & Bailin, 2021, 33-34).

Along the same lines, Browne (2021, 221) also explains that a question-based critical thinking process is one of the best methods of approaching critical thinking as it becomes a community-building process, rather than a cross-examination and judgmental undertaking. He terms this process collaborative critical thinking, a communal endeavour, based on mutual respect and friendliness, which fosters the development of constructive relationships, and not a struggle or a competition in which there are winners or losers. Thus, the critical thinking process becomes a dialogue in itself, an active endeavour to identify the strong and weak points of a dialogue or conversation, an active engagement on the part of the critical thinker with the subject matter presented, and the end result is a stronger understanding of the principles of reasoning that enriches and develops the whole community.

d) <u>Narrative analysis</u>

Narratives perform a vital function in human communication as they respond to the human need to understand one's environment and experiences, to make sense of the world. To this end, narratives expound causal, temporal and spatial relationships that order events and confer meaning and are, therefore, explanatory, correlational and organizational in nature.

Fisher (1987 64-65) explains that the narrative paradigm which underpins human communication, interaction and societal development revolves around five presuppositions:

1. Humans are first and foremost storytellers;
2. Human decision-making and communication are focused on "good reasons" which vary according to the situation, the genre, the medium of communication, the culture, etc.;
3. Good reasons are dependent on history, biography, culture, the characters involved;
4. Narratives are assigned varying degrees of rationality based on narrative probability (the coherence of a story) and narrative fidelity (if the stories ring true to the audience's life experiences);
5. Narratives construct the human world and people choose among various narratives as they construct and reconstruct the world they live in. Good reasons, along with symbols, are the "communicative expressions of social reality."

Therefore, narratives rely on good reasons to be convincing and to gain traction with audiences. Evaluating these good reasons is one of the most important tasks of the critical thinker, as well as decoding the symbols and the cultural references they employ.

Halverston et al (2011, 12) posit that narratives are responsible for giving meaning to language, of creating word patterns and phrases which are meaningful and can be decoded appropriately by the audiences. Madisson & Ventsel (2021, 22) further explain that telling a story means that experience is segmented into concrete units which are afterwards ordered in a definitive and meaningful manner, by creating temporal and causal relations. However, this process is based on subjective interpretation and, consequently, one cannot speak of true stories, but only of interpretations which reflect particular interests, values, beliefs, understandings, etc. Colley (2017, 4) defines narratives as "temporally, spatially and causally connected sequence of events, selected and evaluated as meaningful for a particular audience".

Narratives provide a clearly understandable structure and predictability, which is why they are often employed as the main means of aiding the audiences in comprehending and managing crises, conflicts and other types of threatening events. Gottschall (2019, 47) states that narratives reflect but also exercise the human mind's associative processes for the difficult situations that it may be confronted with and that it needs to untangle. Halverson et al. (2011, 14) present narratives as "a coherent system of interrelated and sequentially organized stories that share a common rhetorical desire to resolve a conflict by establishing audience expectations according to the known trajectories of its literary and rhetorical form", which means that narratives operate according to set rules regarding what is plausible and how event could unfold as dictated by the narrative patterns specific to a culture or history.

Holmstrom (2016: 120) also emphasises that narratives are especially important when human beings are confronted with crisis situations, in which information is scarce and which create a void of comprehension. In their attempt to make sense of unfolding events, people are more willing to accept any story that makes some sense, regardless of its merits and validity because "facts alone cannot ease the feeling of being lost intellectually. Narratives answer the basic human need for structure and predictability. If one side fails to provide a meaningful narrative, others will fill the void" (Holmstrom ,2016, 120).

Therefore, it is important for the critical thinker to analyse the ways in which the connections, regardless of their type, are created and how the messages are tailored for specific audiences. More precisely, the critical thinker needs to examine what is being stated in the narrative, but even more importantly, what details are omitted from the narrative, and to what end. This could be done by reviewing narratives regarding the same events from multiple sources, which have different audiences, and parallels could be construed with respect to the missing or altered information in each of the narrative variants.

Colley (2017, 4) explains that narratives have a clear linear structure, with a beginning, a middle and an end, which are socially constructed and involve actors, setting and plot. The narrative is based on a past, which leads to the present and presents a possible future. And in line with Campbell's myth structure, the narrative gains strength when it is centered around a resolution of conflict, meaning that it starts with an initial disruption, presents and explains the necessary steps to solve the issue and then presents the order restored, thus offering a satisfactory resolution. Miskimmon et al. (2013, 7-10) take over the classical structure of the narrative and apply it to international relations in what they call strategic narratives. Miskimmon et al. (2013) explain that strategic narratives are "representations of a sequence of events and identities, a communicative tool through which political actors – usually elites – attempt to give determined meaning to the past, present and future in order to achieve political objectives" (2013, 7). Tatham (2010, 27) insists on the fact that strategic narratives are tailored and targeted to a specific audience "A thematic and sequenced account that conveys meaning from authors to participants about specific events". This means that strategic narratives can be employed to shape the understanding and interactions in international relations as well. To this end, they can mold policies, determine strategic advantages, project desired organisational images, induce certain reactions to crises and conflicts, build expectations with respect to certain actors or situations, produce predictions about future courses of action, etc.

Strategic narratives also consist of: actors, events (plot and time), and setting (including space). The actors present their own characters in the narratives, they depict the events, which could be historical or contemporary, but always based on chronology and causality and meant to support and promote the actors' interests and values. The actors and the events are represented against the backdrop of the setting or the context which involves spatial representations as well as other types of contexts, be they legislative, diplomatic, historical, economic, etc. The setting is vital as it determines the actors' actions and affects the ways in which events unfold in the audience's perceptions. Being in control of the setting means that an actor has more chances of promoting their narrative to the audience and of getting the audience to interpret the respective narrative in the desired vein.

Given all these aspects regarding the ways in which narratives are constructed and promoted and to which end, we believe it is necessary to develop a toolkit for the critical thinker to use in examining them as well. We will continue in the inquiry vein, as we believe it is the most likely to produce the desired interaction with the narratives and their proponents, while at the same time allowing the critical thinker to maintain a detached, yet friendly perspective on them.

- What is the narrative about? What crisis, event, situation is at its core?
- Who are the actors involved?
- What are those actors' interests?
- What events are presented in the narrative?
- What events are not presented in the narrative?
- What causal relationships are constructed? Is there indeed a causal relationship between those events?
- What temporal sequences are indicated? Is that truly the temporal sequence as other sources certify?
- What historical contextual elements are relevant for the narrative? What historical contextual clues does the narrative employ? What historical contextual clues are omitted?
- What spatial setting is included in the narrative? How has the respective spatial setting evolved in time?
- What legislative and/or diplomatic setting is employed in the narrative? Are any relevant legislative and/or diplomatic aspects omitted?

These questions could assist the critical thinker in bringing to light the ways in which the narratives are constructed, the reasons for the inclusion of certain elements as well as for the exclusion of others, and throw light on the involved actors' interests and the extent to which they could be shaping the narratives to serve those interests, possibly to the detriment of others.

Exercising critical thinking

As previously mentioned, in addition to cognitive biases that can occur in the process of interpreting digital content, logical fallacies can also mislead us. These are errors in reasoning that can make an argument invalid or inaccurate. Below are two practical applications of critical thinking for untangling logical fallacies.

The Nirvana Logical Fallacy

For example, the Nirvana Logical Fallacy [48]refers to the tendency to judge a thing or an action in maximal terms, black or white, either/or: either something is perfectly useful, or it is totally useless. Through this cognitive mechanism, a concrete situation in real life is evaluated by comparison with an ideal situation, considered perfect, but hardly plausible. A false dichotomy is thus created between an implausible option, but presented favorably, and a plausible one, illustrated from an unfavorable perspective. In this context, the choice is not between two solutions that are equal from the point of view of plausibility, but between a realistic solution, on the one hand, and an ideal one which, precisely because it is ideal, will never be encountered in reality.

In the case of this error, a false dichotomy intervenes, a choice between a realistic solution, on the one hand, and an ideal one which, being ideal is somehow better. Thus, the Nirvana error can lead to erroneous or dangerous decisions. By pursuing a perfect solution, we can ignore a useful solution; by aiming to completely solve a problem, we may fail to at least improve a situation. Therefore, instead of aiming for the perfect solution, we can aim for a better solution (small improvements will lead to bigger changes in the long run).

2. The "slippery slope" argument

The "slippery slope" argument is one of the easiest logical fallacies to spot. It is called the "slippery slope" because it passes, *ex-abrupto*, without a logical path, from one statement to another, trying to tie the second closely to the first and, thus, to authenticate it. In a slippery slope argument, an action is rejected because, with little or no evidence, it is insisted that it will lead to a chain reaction that will result in an undesirable goal or end. The slippery slope involves accepting a sequence of events without direct evidence that these events will happen.

Often this fallacy appeals to people's emotions or fears. The problem with this reasoning is that it avoids engaging with the issue at hand and instead directs attention to hypothetical extremes.

[48] The Nirvana fallacy : when perfectionism leads to unrealistic solutions https://nesslabs.com/nirvana-fallacy

This principle is a pseudo-argument that uses a fear-mongering technique and can induce a moral panic.

This type of argument can have the following structure:

"Premise A leads to consequence B, which leads to C, which leads to D, and so on. The final result is then used to state why the initial premise ('A') is wrong'

If we allow A to happen, then Z will happen, and therefore A should not happen. For example: "If we provide free healthcare, then where do we stop? Soon people will be asking for free cars, free cell phones, free food and free everything. The more people get free stuff, the less they will work, which will ultimately lead to economic crises."

Although people may unintentionally use fallacious "slippery slope" arguments, either during discussions or as part of their own reasoning process, these fallacious arguments are often intentionally used as rhetorical techniques because they can be quite persuasive when implemented correctly. Consequently, slippery slopes are often combined with emotion appeals, usually with the aim of appealing to negative emotions such as fear or hatred, but sometimes with the aim of appealing to positive emotions such as hope or compassion.

There are various approaches you can take when responding to a slippery slope argument. For example, slippery slope arguments often omit important events that connect between the start and end points of the slope, and highlighting these can help illustrate problems in this argument. Also, the more disconnected and distant the pieces of the slope are from each other, the less valid the argument. Slippery slope arguments can be either valid or wrong; their validity depends on a number of factors, such as the likelihood that the initial event in question will lead to the intended end result and the wording used to convey that likelihood.

3. The "straw man" argument

The "strawman" argument[49] is an informal logical fallacy that relies on distorting an opponent's claim so that it becomes easier to refute. He thus places himself in the same rhetorical zone as the "slippery slope" argument.

For example, let's say that a doctor - X - would say: "to avoid getting sick with SARS-CoV-2, it's good to avoid crowded areas". His opponent, Y, would use the "straw man" argument like this: "so, according to Dr. X, the only solution would be for everyone to stay indoors", continuing - "they are using the pandemic to keep us in detention in our own homes". The logical fallacy is easy to spot just by going back to Dr. X's statement, which was not talking about

[49] Strawman Fallacy https://www.logicallyfallacious.com/logicalfallacies/Strawman-Fallacy

detention, but only about avoiding crowding; doctor X therefore makes a statement that Y does not logically debunk, but distorts it so that it becomes repulsive and difficult for the public to accept.

In this case, one attacks a position that the opponent does not really hold. Thus there is an oversimplification, taken out of context or exaggerated of a perspective. This fallacy in argumentation is meant to distract from the real issue being discussed and is not a logically valid argument. This pseudo-argument is particularly common in political debates and discussions of controversial topics. The basic structure of the argument consists of:

"Person A holds point X, Person B creates a distorted version of point X ("straw man"), and then Person B attacks this distorted version to refute Person A's original claim"

Additional resources:

- ✓ 5 Examples of The Nirvana Fallacy https://simplicable.com/new/nirvana-fallacy
- ✓ Nirvana Fallacy https://www.logicallyfallacious.com/logicalfallacies/Nirvana-Fallacy
- ✓ Straw Man Fallacy Examples https://examples.yourdictionary.com/straw-man-fallacy-examples.html
- ✓ Bad Arguments and How to Avoid Them https://fs.blog/bad-arguments/

Measuring critical thinking

Haines and Stein (2021) explain that the Critical thinking Assessment Test (CAT) is based upon one such contemporary, inclusive, skill-based approach, and the skills that it measures have proven reliable indicators to assess students' critical thinking competencies. These skills are:

1. Evaluating information

- Separate factual information from inferences
- Interpret numerical relationships in graphs
- Understand the limitations of correlational data
- Evaluate evidence and identify inappropriate conclusions

2. Creative thinking

- Identify alternative interpretations for data or observations
- Identify new information that might support or contradict a hypothesis
- Explain how new information can change a problem

3. Learning and problem solving

- Separate relevant information from irrelevant information
- Integrate information to solve problems

- Learn and apply new information
- Use mathematical skills to solve real-world problems

4. Communication

- Communicate ideas effectively (Haynes and Stein 2021 237)

References:

1. Bailin, S. and Battersby, M. (2021) "Inquiry: Teaching for Reasoned Judgment" in *Critical Thinking and Reasoning Theory, Development, Instruction, and Assessment.* Daniel Fasko, Jr. and Frank Fair (eds.), pp.31-46.
2. Browne, Neil M. & Keeley, Stuart M. (2007) Asking the right questions. A Guide to Critical Thinking, Eighth Edition, Pearson Prentice Hall.
3. Browne, Neil M. (2021) "Commentary: Critical Thinking – Effusively Touted, but so Rarely Pursued"**,** in *Critical Thinking and Reasoning Theory, Development, Instruction, and Assessment.* Daniel Fasko, Jr. and Frank Fair (eds.), pp. 209-227.
4. Campbell, Joseph, & Moyers, Bill (1991). *Power of Myth*. Bantam Doubleday Dell Publishing Group.
5. Colley, T. (2017). Is Britain a force for good? Investigating British citizens' narrative understanding of war. *Defence studies*, *17*(1), 1-22.
6. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the Digital Education Action Plan, 2018, available at EUR-Lex - 52018DC0022 - EN - EUR-Lex (europa.eu)
7. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on achieving the European Education Area by 2025, 2020, available at EUR-Lex - 52020DC0625 - EN - EUR-Lex (europa.eu)
8. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions A NEW SKILLS AGENDA FOR EUROPE, 2016, available at EUR-Lex - 52016DC0381 - EN - EUR-Lex (europa.eu)
9. Facione, P. A. (1990a). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations prepared for the Committee on Pre-College Philosophy of the American Philosophical Association. ERIC Document ED315423.
10. Fisher, Walter R. (1987) *Human Communication as Narration*, University of South Carolina Press.
11. Gottschall, Jonathan (2013). *The Storytelling Animal: How Stories Make Us Human*, Boston, New York: Mariner Books.
12. Govier, T. (1989) Critical Thinking as Argument Analysis?, *Argumentation* 3: 115-126.
13. Halverson, J., Corman, S., & Goodall, H. L. (2011). *Master narratives of Islamist extremism*. Springer.
14. Haynes, A. and Stein, B. (2021) "Observations from a Long-term Effort to Assess and Improve Critical Thinking" in *Critical Thinking and Reasoning Theory, Development, Instruction, and Assessment*. Daniel Fasko, Jr. and Frank Fair (eds.), pp.231-254.
15. Holmstrom, Miranda (2016). The Narrative and Social Media, available at https://www.stratcomcoe.org/mirandaholmstrom-narrative-and-social-media
16. Hunter, David A. (2009) *A Practical Guide to Critical Thinking. Deciding What to Do and Believe*, Wiley, John Wiley & Sons Inc.
17. Kahane, H. (1989) The proper subject Matter for Critical Thinking Courses, *Argumentation* 3: 141-147.
18. Kinneavy, J. L. (1986). Kairos: A neglected concept in classical rhetoric. *Rhetoric and praxis: The contribution of classical rhetoric to practical reasoning*, 79-105.

19. Kinneavy, J. L., & Eskin, C. R. (2000). Kairos in Aristotle's rhetoric. *Written communication*, *17*(3), 432-444.
20. Levintin, D. (2017) *Weaponized Lies: how to think critically in the post-truth era*, Penguin Random House.
21. Lotman, Y. M. (1990). Universe of the Mind. *A semiotic theory of culture*, 20-35.
22. Madisson, M. L., & Ventsel, A. (2021). *Strategic conspiracy narratives: A semiotic approach*. Routledge.
23. Miskimmon, A., O`Loughlin, B. and Roselle, L. (2013) *Strategic Narrative: A new means to understand soft power*, New York, London: Routledge.
24. Miskimmon, A., O`Loughlin, B. and Roselle, L. (eds.) (2017). Forging the world: strategic narratives and international relations, Ann Arbor: University of Michigan
25. Press.
26. Paul, R. and Elder, L. (2004) *The Thinker's Guide for Conscientious citizens on How to Detect Media Bias & Propaganda in National and World News*, The Foundation for Critical Thinking.
27. Paul, R. and Elder, L. (2020) *The Miniature Guide to Critical Thinking Concepts and Tools* Eighth Edition, Rowman & Littlefield.
28. Siegel, H. (1989). The rationality of science, critical thinking, and science education. *Synthese*, *80*(1), 9-41.
29. Snyder, T. (2017) *On Tyranny. Twenty Lessons from the Twentieth Century*, Tim Duggan Books, New York.
30. Tatham, S. (2010). Understanding strategic communication: Towards A definition. *Strategic Communications for Combating Terrorism. Ankara: Centre for Excellence Defence against Terrorism*, 17-37.
31. Walton, D. (1989) Dialogue Theory for Critical Thinking, *Argumentation* 3: 169-184

3.3 Media Literacy

Irena Chiru

Abstract

The current section acknowledges the various attempts of theorising the competencies needed for an informed and aware media/ news consumption, while trying to bring into discussion the elements that forms media literacy in a media interconnected society as it is the society we are presently experiencing. Within the broader scope of the DOMINOES Handbook, this section aims to present and analyse the most recent developments in developing media competencies. It stems from the premise that with the multiple problems of hate speech, cyberbullying, hacked YouTube content or fake news we need to better understand and manage the media environment. In addition to exploring the alternative approaches of media literacy as an over-arching concept for information literacy, digital literacy, and fake news literacy, the section will identify, structure and investigate both the challenges and the inspiring practices in understanding the role of media in society and promoting a critical approach to messages constructed by media. Given the specificity of the DOMINOES Handbook, the focus will be mainly on the actions assumed at the European level and, out of all media contents, on news.

Main research questions addressed

- How has media literacy evolved in the last years? How has the intensive debate on disinformation and fake news shaped the "media literacy" initiatives?
- Which are the main media competencies needed in the current information ecosystem?
- What are the good practices in the field?

Media literacy represents a widely invoked solution facing the dangers of misinformation, disinformation and the use of information to cause harm. Also found as "information literacy" and/or in close relation with "media education", media literacy can be simply defined as the ability to:

- decode and understand media messages (including the political and economic ecosystems in which they are produced and exist);
- assess the influence of those messages on human beliefs, feelings, and behaviours;
- repost/ create mediatised content thoughtfully and conscientiously.

Regardless of the field in which it is considered (e.g. financial, digital, computer or cultural), "literacy" is understood as a desirable state or "something we seek to achieve" (Leaning, 2017, 30), in which one has or aims for a level of understanding beyond simple competence based on cognitive skills and reasoning. The existing literature displays a large set of alternative definitions (e.g. Aufderheide, 1993, Potter, 2010, Polanco-Levicán, Salvo-Garrido, 2022) that vary from the ability to access, understand and produce media content in a variety of contexts to an informed and skillful application of literacy skills to media and technology messages. The literature also offers a full panoply of concepts and definitions related or subordinated to media literacy, as well as their evaluation and comparison, however arriving at a universally applicable and practical model is impossible and would be unworkable. Therefore, the current section acknowledges the various attempts to theorise the competences needed for an informed and aware media and news consumption, while trying to bring into discussion the elements that form media literacy in a media interconnected society as is the society we are presently living in.

Media literacy definitions

- “A critical-thinking skill that enables audiences to decipher the information that they receive through the channels of mass communications and empowers them to develop independent judgments about media content” - Silverblatt and Eliceiri (1997)
- “The ability to effectively and efficiently comprehend and use any form of mediated communication” – Baran (2013)
- “[media literacy] includes all technical, cognitive, social, civic and creative capacities that allow a citizen to access, have a critical understanding of the media and interact with it” (*EU Media Literacy Expert Group*).

Although seldom incorrectly and flexibly used as a replacement for ”information literacy” or as an obsolete formula for ”digital literacy”, ”media literacy” mirrors the result of the media convergence – that is the merging of electronic media (mass communication) and digital media (multimedia communication). Therefore, the formula embraced by the current section - media literacy - is to be considered as the appropriate concept leaving behind or including former or subordinated types of literacy:

- classic literacy (reading-writing-understanding) was dominant for centuries and corresponded to the process of reading and writing, and in which primary schooling has played an essential role;
- audiovisual literacy, which relates to electronic media such as film and television, focuses on image, and sequential images. It is the beginning of different educational initiatives early engaged but not sufficiently supported by a real policy;
- digital literacy or information literacy stems from computer and digital media, which brought about the necessity to learn new skills. This is a very recent concept, and is often used synonymously to refer to the technical skills required for modern digital tools which occurs in the advanced stages of development of information society.

The recent conceptualising and measuring initiatives at the European level (Celot, 2012) identified two dimensions within media literacy. The first one is derived from an individual’s ability to utilise the media, Individual Competencies, defined as:

(a) Use – an individual technical skill;

(b) Critical Understanding competency – fluency in comprehension and interpretation and

(c) Communicative – the ability to establish relationships through the media and the other informed by contextual and environmental factors.

The second dimension, Environmental Factors, is defined as a set of contextual factors that facilitate or hinder the development of the Individual competencies including the following areas: (a) Media education, (b) Media Policy, (c) Media Availability, (d) Roles of the Media Industry and Civil society.

From a more nuanced perspective, media literacy is characterised by eight fundamental characteristics (Silverblatt, 2008):

1. A critical thinking skill enabling audience members to develop independent judgments about media content;
2. An understanding of the process of mass communication;
3. An awareness of the impact of media on the individual and society;
4. Strategies for analysing and discussing media messages;
5. An understanding of media content as a text that provides insight into our culture and our lives;
6. The ability to enjoy, understand, and appreciate media content;
7. Development of effective and responsible production skills;
8. An understanding of the ethical and moral obligations of media practitioners.

Media Literacy Fields of Application

(*A European approach to media literacy in the digital environment*, 2007)

This reference material on media literacy published by the European Commission places the primary focus of development on the following three fields:

- “online content - empowering users with tools to critically assess online content - extending digital creativity and production skills and encouraging awareness of copyright issues - ensuring that the benefits of the information society can be enjoyed by everyone, including people who are disadvantaged due to limited resources or education, age, gender, ethnicity, people with disabilities (e-Accessibility) as well as those living in less fortunate areas (all these are encompassed under e-Inclusion);
- raising awareness about how search engines work (prioritisation of answers, etc.) and learning to better use search engines;
- commercial communication - giving young audiences tools to develop a critical approach to commercial communication, enabling them to make informed choices - encouraging public/private financing in this area with adequate transparency;
- audiovisual works - providing, notably to young European audiences, better awareness and knowledge about our film heritage and increasing interest in these films and in recent European films - promoting the acquisition of audiovisual media production and

creativity skills - understanding the importance of copyright, from the perspective of both consumers and creators of content".

Hence, media-literate people develop critical thinking skills enabling "to develop independent judgments about media and media content" (Baran, 2014) and to have an awareness of the impact of media on the individual and society. News media literacy has proved to be effective in modelling several psychological or behavioral variables, such as event knowledge (Vraga et al., 2011), political efficacy (Semetko & Valkenburg, 1998), and conspiracy theory endorsement (Craft et al., 2017). Other surveys have illustrated a positive association between news media literacy and current events knowledge (e.g., Ashley et al., 2017). Moreover, news media literacy was also found to facilitate individuals' skeptical attitudes toward news content (e.g. Maksl et al., 2015; Vraga et al., 2015). Scholars have suggested that an important feature of news media literacy pertains to the ability of general inquiry and critical thinking; hence, highly media literate individuals usually are skeptical of the media content due to their familiarization with media practice routines, and better understanding of the news production and dissemination environment (Vraga et al., 2015).

Present state of affairs – what is new?

In line with the changes experienced by the media (from traditional to digital) and the growing quantity of information we daily produce and consume, in the last decade media and information literacy has been undergoing a 'great turn' (Hargreaves et al. 2010), meaning a period of rapid transition and change in practices. This switch can be easily attributed to the emergence and unprecedented use of digital technologies and to the irreversible and incontrollable claim they have on the economic, political and societal arenas. By comparison to the moment when this significant turn in media education was initially theorised, the effect has grown exponentially from the political weaponisation of misinformation and its use during presidential elections, to large dissemination of conspiracy theories, to viral COVID-19 infodemic during pandemic, and "organized social media manipulation" or "industrialized disinformation" (Bradshaw et al. 2020).

Accordingly, the specificity of the new digital media (Internet, mobile) is challenging the traditional approach to the media literacy and education. Due to their interactive specificity, digital media raises additional problems: not only the risk of passive consumption and a lack of critical thinking (as in one-way media), but a reshaping of the modalities in which people interact with them and the difficulties encountered in regulating the development of media content and in controlling its effects. The new roles – from receivers of messages as with traditional media to acting effortlessly as creators and producers - added new challenges and made media literacy education more complicated. So, in the current age of media literacy, to the competencies of accessing, analysing, evaluating, and creating media messages across a variety of contexts, we must add the creative and playful forms of multimodal media content production, as well as abilities to reflect on one's communication behaviour, to act and participate in society (Cannon et al., 2018).

In the context of the accelerated misuse of information for deception in the public arena, media literacy has been redefined as "fake news literacy" that is as "the individuals' ability to discern fake news from real news" (Huber et al. 2022). However, existing results have not so far been able to significantly relate general media literacy, news literacy or digital literacy to the accurate identification of fake information. Surely, fostering people's general media news literacy skills can be helpful in addressing to disinformation, but only the developing of specific fake news literacy skills can make a difference since only those will enable individuals to engage in corrective actions.

Looking at the initiatives dedicated to fostering media literacy in the last 20 years, we cannot help but notice the high level of consensus about the need for public policy to give special attention to the promotion of media literacy. Gradually, the initial focus on introducing ICT skills into the education system as the main means of media education and media literacy was extended by countries appointing specific departments (ministerial departments, pubic companies or other) to promote media literacy skills among citizens, and the launch of campaigns promoting media literacy. Most countries have modified their curricula to include digital and media skills: the United Kingdom, Spain, France, Finland, Italy and Portugal but only in certain countries is the promotion of skills related to digital literacy extended to the mass media and general communication, that is, to media literacy. Such countries include Germany and Finland. Nevertheless, the dominant trend is that there is "no complete convergence between the digital and media curriculum, meaning that problems that could be resolved with an integrated framework still remain without solution" (*Study of the Current*). In addition, several studies that were conducted in the last decade showed that the meanings of media literacy and the formal (in schools) or informal practices associated with media education significantly vary from one state member to another (e.g. EMEDUS, 2014).

According to the report *Mapping of media literacy practices and actions in EU-28 9* (European Audiovisual Observatory, Strasbourg 2016), the main media literacy skills addressed by European initiatives are: "creativity, critical thinking, intercultural dialogue, media use and participation and interaction". Results show a rather disproportionate implication of different European countries in joint European or international endeavours. Most of these projects are carried out by cross-sector collaboration, civil society, public authorities and have as main target the group of the teens/ older students.

With the advent of social media and their insufficient content regulation facilitating a dynamic environment where mis- and disinformation are spread, the European Commission approach to media literacy (concepts and practices) has been gaining ground. For example, the revised *Audiovisual Media Services Directive* strengthens the role of media literacy also by "requiring Member States to promote measures that develop media literacy skills" (adopted by the Council in 2018). The Directive obligates "video-sharing platforms to provide effective media literacy measures and tools, this being a crucial requirement due to the central role such platforms play in giving access to audiovisual content". In addition, platforms are also required "to raise users' awareness of these measures and tools". Furthermore, the European Commission supports other initiatives such as the *Media literacy expert group* which aims to identify, document and extend good practices in the field of media literacy, facilitate networking between different stakeholders and explore ways of coordinating EU policies, support programmes and media literacy initiatives or the *European Media Literacy Week* and *European Media Literacy Awards* (https://digital-strategy.ec.europa.eu/en/policies/media-literacy).

Main challenges

Challenge 1: Media literacy is not the only solution to the problem of disinformation and should stop being hailed as a silver-bullet solution

Although seldom considered as a one-shot campaign, media literacy must be reconsidered as a long-term solution requiring thought-through pedagogical strategies and years of teaching. In this sense, S. Livingstone defines it as a moving target (https://blogs.lse.ac.uk/medialse/2018/10/25/media-literacy-what-are-the-challenges-and-how-can-we-move-towards-a-solution/). As society becomes more dependent on the media and the media are becoming more complex, fast-changing, commercial and globalized, media literacy strategies require sustained attention, resources and commitment – to education, to curriculum development, to teacher training, to research and evaluation. Hence, although media literacy tends to be seen by policy makers as an easy win, in fact it is incredibly complicated. There is so much about the online world that is illegible and constantly changing, and it is crucial to avoid burdening the citizen with the responsibility of understanding the incomprehensible.

For example, one of the constant sources of dispute conferring complexity to the topic is the relation between developing media competencies and promoting regulations for reducing/ eliminating the online harmful content. Because many regulations will be less effective without an accompanying level of education and awareness, media literacy is an essential partner to regulation in terms of improving the public's ability to navigate the online environment.

Challenge 2: The vague status of media education as a curriculum subject

Over the last decades, media education has significantly been embraced in formal and informal working formats by many nonprofit organisations and governments. However, despite the numerous initiatives, media education has a vague status. For example, there is considerable uncertainty about whether media education should be regarded as a separate curriculum subject, or just be considered a part in existing other study units. Where existing, the formal practices associated to media literacy in different schools are highly dependent on the circumstances provided by different national contexts (Buckingham and Domaille, 2001). It appears most frequently as "a 'pervading' element of the curriculum for mother tongue language or social studies without being a well-established and clearly defined area of study" (Buckingham and Domaille, 2001). Hence, the dominant trend is that there is no complete convergence between the digital and media curriculum, meaning that problems that could be resolved with an integrated framework still remain without solution.

The aim of digital literacy is to help people to become active and conscious citizens of the information society (Rivoltella, 2006). In school, this does not mean to make place for a new subject, but to develop a cross-curricular approach so that students have the chance to learn in a digital environment and teachers to adopt media and communication as a teaching style.

Interactivity and user content generation could be the new methodological perspectives of this new paradigm.

Challenge 3 Developing media competencies require a tailored-made approach

Different groups of people require different media literacy interventions at different points on their learning journeys. If in the traditional approach of media literacy it was mainly addressed to children and youth as part of their initiating educational processes, current perspectives must be tailored as to meet different needs. What is the impact of the age differences in the ability to identify fake news? To what extend do analytical reasoning, affect and news consumption frequency effect this ability?

For example, recent research has identified older adults as a demographic group in particular vulnerable in front of fake news. Several surveys conducted during the COVID 19 pandemic when presumably social technologies were used more than usual for critical social interaction, showed that the more elderly older adults showed a reduced ability to detect fake news, not just on COVID but on any other topic, and that "decreased ability was associated with levels of analytical reasoning, affect and news consumption frequency" (Pehlivanoglu et al. 2022). In particular, the individuals age 70 or older form a high-risk population with high stakes for engaging in "'shallow' information processing, including not looking as closely at information or paying attention to details". This results are valid even when discussing the problem outside the pandemic framing - older adults are more likely than all other age groups to incorrectly think that news encountered in their Facebook Newsfeed are filtered by professional editors and journalists, which might make them more likely to trust and share misinformation encountered in that environment (Fletcher et al, 2020). One of the main conclusions of these surveys is that besides family and friends technology companies along with support older adults' use of digital media, by providing resources like technology tutoring, loaner devices, or educational content about new platforms and online safety (Moore and Hancock, 2022).

Children can be targets and objects of disinformation, spreaders or creators, but can also act as opponents of disinformation by actively seeking to counter falsehoods. When it comes to youngsters, the data collected in several European countries and Quebec illustrate that 12 to 18-year-olds develop numerous and shared uses in fundamental domains such as ethics or social issues of IT, but their appropriation remains incomplete, mostly in information and creative activities (Bevort and Breda, 2008). The study also highlights how young people appropriate digital media and how their practices differ within different contexts of use (at school and at home, for example). The landscape depicted is certainly more optimistic than usually considered: youngsters are more critic and conscious than we might expect. On the contrary, educational institutions, that is, essentially the school, but also associative educational spaces and media do not seem to have measured the importance that the new media have acquired in the daily lives of young people remaining unable to act so that educational challenges coming from these media could be accepted finding their answers. However, additional research is needed in the field as for researchers and policymakers to get a clear and comprehensive picture of how susceptible children are to disinformation and how it affects their development, well-being and rights (Howard et al., 2021).

Inspiring practices, projects, interventions in the field

Example 1 Learning from Finland as top 1 media literate European country

Finland (1st), Denmark (2nd), Estonia (3rd), Sweden (4th) and Ireland (5th) are at the top of the ranking of the *Media Literacy Index 2021*. These are countries that have the highest potential to address the negative impact of fake news and misinformation due to the quality of their education, free media and high level of trust among people. Taking a closer look at Finland, that has remained no 1 among the 35 European countries during the last years, several characteristics of media literacy projects and useful conclusions can be extracted:

- media education is present throughout Finland's education curriculum. Starting in day-care, media education continues via lifelong learning aiming to reach out to everyone.
- recognising disinformation is considered to be important, but it is only ”a small part of media education” (Pekkala, https://finland.fi/life-society/educated-decisions-finnish-media-literacy-deters-disinformation/). According to the Finnish governmental strategy, media literacy by itself is not the end goal, but it is a means combining both the technical skill to use media and the ability to understand it with the scope of becoming a good citizen and thus contributing to a stable democracy and a healthy society.
- media education in Finland takes an approach that includes the whole of society. Many different civic organisations take part in developing and enacting learning programmes, including schools, libraries, government departments, universities and NGOs. To this, various actions, campaigns and training formats are implemented.

Example 2 eTwinning[50]

eTwinning represents a learning community for teachers in Europe, which published in 2021 the book ‘Teaching Media Literacy and fighting Disinformation with eTwinning’. This book aims to inspire and support teachers and pupils of all ages by exploring the multiple aspects of this topic, illustrating examples of eTwinning projects, and offering resources and activities. It explains the concept of disinformation, looks at how young people engage with media, showcases outstanding eTwinning projects on media literacy and disinformation, and gives examples of tools and resources, as well as of classroom activities for developing media literacy (book2021_etwinning_interactif-1.pdf).

[50] eTwinning is the community for schools in Europe. eTwinning offers a platform for staff (teachers, head teachers, librarians, etc.), working in a school in one of the European countries involved, to communicate, collaborate, develop projects, share and, in short, feel and be part of the most exciting learning community in Europe, https://school-education.ec.europa.eu/en/etwinning.

Example 3 SPreaD

By developing a toolkit on the management of digital literacy projects SPreaD aims at disseminating digital literacy all over Europe and to raise awareness on this important topic. The SPreaD toolkit gives useful tips regarding the development, coordination and financing of large scaled digital literacy projects (http://www.spread-project.eu/).

Example 4 Educational Toolkit for the Development of Social Media Literacy (ETDSML)

It is an innovative educational project aiming to develop social media literacy. It created an educational toolkit for teachers teaching social media literacy in school including a curriculum for the development of social media literacy in schools, a course support for the development of social media literacy in schools, a methodological guide for teaching social media literacy, a mobile application for implementing social media literacy in school and a collection of learning scenarios including the use of social media tools (https://socialmedialiteracy.eu/).

The variety of definitional elements, the diversity of interpretations of widely quoted definitions, and the frequent citing of alternative sources for the same idea leads to the conclusion that scholars who write about media literacy exhibit considerable variety in their meanings for the term. It appears that everyone who writes about media literacy has a different perspective on what it is or what it should be, unless we keep our focus at the most general level of meaning. This raises the question about how this sharing of meaning only at the most general level benefits or limits the development of media literacy as a scholarly field.

As shown above, the definition of media literacy is very broad and, although it has recently included digital competencies, it cannot be reduced to the use of the Internet or computers. More than the technological ability to use a computer, media literacy focuses on users' capacity to receive and evaluate information in the digital environment and to use this information in an effective and responsible way. Therefore, becoming and remaining digitally and media literate is a continuous process that requires exercise in non-formal and formal working formats, from early stages of education to special programs for elders.

Facing the challenges of our multimedia society, characterised by interactivity, portability and connectivity, we have recently witnessed an intensification of initiatives aiming at fostering media competencies as a defensive shield against the individual and social perils of disinformation. These involve a wide range of actors - governmental, educational, NGOs, private ones as and no single organisation or sector can be expected to achieve this range of media literacy support on their own. But there are so many initiatives that an attempt to try to find the “right one” may be overwhelming. To ensure efficiency future endeavours should also cover the following questions: How can we better structure and push forward the already implemented solutions? How can we better make all these project be convergent while avoiding redundancy?

Although frequently invoked as an all-in-one solution in fighting disinformation, along with critical thinking, media literacy must be reconceptualised as to specifically address the challenges of disinformation and of the new information age. Also, a continuous and systematic empirical data-driven approach is needed as the landscape of information is so quickly changing. At a global scale, the deficiencies in digital media literacy have been considered as one of the significant factors explaining the widespread belief in online misinformation; in the recent years, this observation determined changes in education policy and the design of technology platforms. However, in order to be properly oriented, polices must be systematically informed by rigorous evidence regarding the relationship between digital media literacy and people's ability to distinguish between information and disinformation, low- and high-quality news online (Guess et al., 2020).

A very recent swift in designing media literacy projects has been given by the focus on individual responsibility in combating the threat of disinformation by becoming more information literate. For the future, such a model, whose focus is the idea of citizenship, seems to be the most wanted and sustainable given current variables in place. Given these challenges, we need to invest more into human-centered solutions focused on improving people's media and information literacy. They not only demonstrate a much deeper and longer-lasting impact, but also may be easier and cheaper to implement than commonly believed. Last but not least, no matter how attractive or adapted to targeted audiences media literacy initiatives are, the media literacy interventions that do not form and cultivate users' motivation to resist such influences are doomed to failure.

References

1. *** (2007). A European approach to media literacy in the digital environment, available at https://www.cedefop.europa.eu/en/news/european-approach-media-literacy-digital-environment, last accessed October 9.
2. *** *(2014). Study on the current trends and Approaches to Media literacy in Europe*, available at https://ec.europa.eu/assets/eac/culture/library/studies/literacy-trends-report_en.pdf, last accessed September 9.
3. *** (2021). Media Literacy Index 2021, available at https://osis.bg/?p=3750&lang=en, last accessed October 3.
4. *** (2021). *Teaching Media Literacy and fighting Disinformation*, available at https://euneighbourseast.eu/news/latest-news/etwinning-releases-new-e-book-on-teaching-media-literacy-and-fighting-disinformation/, last accessed October 9.
5. *** *Educated Decisions: Finish media Literacy deters disinformation*, available at https://finland.fi/life-society/educated-decisions-finnish-media-literacy-deters-disinformation/, last accessed October 9.
6. ***EMEDUS European Media Literacy Education Study REPORT ON FORMAL MEDIA EDUCATION IN EUROPE, 2014, available at https://eavi.eu/wp-content/uploads/2017/02/Media-Education-in-European-Schools-2.pdf, last accessed October 9.
7. Aufderheide, Patricia (1993). *Media Literacy. A Report of the National Leadership Conference on Media Literacy*, https://eric.ed.gov/?id=ED365294, last accessed September 23.
8. Baran, Stanley. J. (2014). *Introduction to Mass Communication: Media Literacy and Culture*, 8th ed. New York: McGraw-Hill.
9. Bevort, Evelyne & Breda, Isabelle (2008). ”Adolescents and the Internet: Media Appropriation and Perspectives on Education. In Pier Cesare Rivoltella (ed.) Digital Literacy: Tools and Methodologies for Information Society, IGI Global.

10. Bradshaw, Samantha, Bailey, Hannah, Howard & Philip N. (2021). *Industrialized Disinformation: 2020 Global Inventory of Organised Social Media Manipulation*. Working Paper, Oxford, UK: Project on Computational Propaganda.
11. Buckingham, David & Domaille, Kate (2001). *Where Are We Going and How Can We Get There?*, available at https://www.researchgate.net/publication/234560105_Where_Are_We_Going_and_How_Can_We_Get_There_General_Findings_from_the_UNESCO_Youth_Media_Education_Survey_2001_Occasional_Paper, last accessed October 9.
12. Cannon, Michelle, Connolly, Steve & Parry, Rebecca (2018). *Media Literacy, Curriculum and The Rights of the Child Authors*, available at https://uobrep.openrepository.com/bitstream/handle/10547/624629/Main%20document%20with%20full%20author%20details%20%26%20affiliations.pdf?sequence=2&isAllowed=y, last accessed October 9.
13. Celot, Paolo (2012). ”EAVI Studies on media literacy in Europe”. In *Stručni rad* / UDK 316.774:8(4), 316.77(4): 303.
14. Craft, Stephanie, Ashley, Seth & Maskl, Adam (2017). ”News media literacy and conspiracy theory endorsement”. In *Communication and the Public*, available at https://www.researchgate.net/profile/Seth-Ashley-4/publication/320219984_News_media_literacy_and_conspiracy_theory_endorsement/links/5dc8f37a458515143500799d/News-media-literacy-and-conspiracy-theory-endorsement.pdf, last accessed October 1.
15. Guess, Andrew M., Lerner, Michael, Lyons, Benjamin, Montgomery, Jacob M., Nyhan, Brendan, Reifler, Jason & Sircar, Neelanjan (2020). ”A digital media literacy intervention increases discernment between mainstream and false news in the United States and Indi”. In *PNAS*, available at https://www.pnas.org/doi/10.1073/pnas.1920498117, last accessed October 15.
16. Fletcher, Richard, Newman, Nic & Schulz Anne (2020). *A Mile Wide, an Inch Deep: Online News and Media Use in the 2019 UK General Election.* Reuters Institute, available at file:///C:/Users/Utilizator/Downloads/Fletcher_News_Use_During_the_Election_FINAL%20(1).pdf, last accessed October 10.
17. Hargreaves, Andy, Lieberman, Ann & Fuller, Michael (2009). ”Introduction: Ten years of change”. In Andy Hargreaves, Ann Lieberman, Michael Fuller (eds.) *Second international handbook of educational change*. London: Springer.
18. Howard, Philip N., Neudert, Lisa-Maria & Prakash, Nayana (2021). *Digital misinformation / disinformation and children*, UNICEF Office of Global Insight and Policy, available at https://www.unicef.org/globalinsight/reports/digital-misinformation-disinformation-and-children, last accessed October 2.
19. Huber, Brigitte, Borah, Porismita & Gil de Zúñiga, Homero (2022). ”Taking corrective action when exposed to fake news: The role of fake news literacy”. In *Journal of Media Literacy Education, 14*(2), 1-14.
20. Maksl, Adam, Ashley, Seth, & Craft, Stephanie (2015). ”Measuring News Media Literacy”. In *Journal of Media Literacy Education*, 6(3), 29-45.
21. Media Appropriation and Perspectives on Education”. In Kristin Klinger (ed.) *Digital Literacy: Tools and Methodologies for Information Society*, IGI Global.
22. Moore, Ryan. C., Hancock & Jeffrey. T. (2020). ”Older Adults, Social Technologies, and the Coronavirus Pandemic: Challenges, Strengths, and Strategies for Support”. In *Social Media + Society*, 6(3).
23. Moore, Ryan. C., Hancock & Jeffrey. T. (2022). ”A digital media literacy intervention for older adults improves resilience to fake news”. In *Scientific Reports*, 12, 6008.
24. *Online media literacy: Across the world, demand for training is going unmet*, available at https://www.ipsos.com/en/online-media-literacy-across-world-demand-training-going-unmet, last accessed October 15.
25. Pehlivanoglu, Didem, Lighthall, Nichole, Lin, Tian & Chi, Kevin J. (2022). ”Aging in an "infodemic": The role of analytical reasoning, affect, and news consumption frequency on news veracity detection”. In *Journal of Experimental Psychology Applied*, 28(3).

26. Polanco-Levicán Karina & Salvo-Garrido Sonia (2022). ”Understanding Social Media Literacy: A Systematic Review of the Concept and Its Competences”. In International Journal of *Environmental Research and Public Health*. 2022 Jul 20, 19(14): 8807.
27. Potter, James (2010). ”The state of media literacy”. In *Journal of Broadcasting & Electronic Media*. 2010, 54(4), 675–696.
28. Rivoltella, Pier Cesare (2008). *Digital Literacy: Tools and Methodologies for Information Society*, IGI Global.
29. Semetko, Holli. A., & Valkenburg, Patti. M. (1998). ”The impact of attentiveness on political efficacy: Evidence from a three-year German panel study”. In *International Journal of Public Opinion Research*, 10(3), 195-210.
30. Silverblatt Art, Eliceiri, Ellen & Eliceri, Ellen (1997). *Dictionary of media literacy* Westport: Greenwood Press.
31. Vraga, Emily. K. (2011). Dealing with dissonance: Responding to an incongruent test result in a new media environment. Cyberpsychology, Behavior and Social Networking, 14, 689–609.
32. Vraga, Emily K., Tully, Melissa, Kotcher, John, E., Smithson, Anne-Bennett & Broeckelman, Melissa Post (2015). ”A Multi-Dimensional Approach to Measuring News Media Literacy”. In *Journal of Media Literacy Education*, 7.

3.2 Debunking, Fact-Checking, Pre-bunking

Ruxandra Buluc, Valentin Stoian-Iordache, Cristina Arribas Mato, Rubén Arcos, Manuel Gertrudix, Alexandra Anghel, Cristina Ivan

Abstract

Debunking and fact-checking have been deemed important methods of countering disinformation. Funds and time have been allotted to implementing technical solutions that could assist in fact-checking online posts and marking them according to their veracity. There are numerous NGOs involved in determining the reliability of online information, and internet platforms such as Google, Facebook, Twitter have departments in charge of checking the content posted in their digital spaces. However, there are limitations to these methods and new ways of countering disinformation focus on inoculation theories and develop means for people to become resilient to disinformation before being exposed to it. This is known as prebunking and is gaining ground. It employs serious games (which will be discussed in detail in section 5) as well as reverse psychology to assist people in developing skills that allow them to spot disinformation and not fall prey to it. The current section focuses on how debunking and fact-checking operate, what their limitations are and what further enhancements prebunking can bring to the fight against disinformation. It is clear from the onset that none of these methods are sufficient on their own. However, joint efforts both before exposure to disinformation and after disinformation becomes viral and is tagged as such, prove to be another effective means of stopping disinformation from spreading and eroding democratic societies, civil trust and cooperation.

Main research questions addressed

- What is debunking?
- What is fact-checking?
- What are the differences between debunking and fact-checking?
- How can fact-checking disinformation help stop its spread?
- What is prebunking?
- What role can reverse psychology play in educating people with respect to disinformation?
- What lessons can be learned from existing examples and practices of effecting debunking and prebunking?

Strategies to combat disinformation have been, until recently, relatively under-researched in comparison to other issues in the broader field of the psychology of belief internalisation. Mahl, Scheffer and Zeng (2022) undertook a synthesis of the available literature on disinformation and on strategies to combat it. They identified several categories of academic literature: those that focus on the willingness to believe or share disinformation, those that looked at the disinformation narratives and those that focused on the strategies to combat disinformation. Recently, effective strategies to combat disinformation have become highly relevant given the wide spread of disinformation on topics such as climate change, vaccination, COVID-19 and the conflict in Ukraine.

Two types of strategies have been identified in the literature on combating mis- and disinformation: (1) debunking, often associated with fact-checking; and (2) pre-bunking. Moreover, the literature also focuses on how psychological factors can be included in the attempts to counter disinformation and more attention has been paid to (3) reverse psychology.

Fact-checking - a short introduction

Fact-checking appeared as a profession after the technological boom that created a 4.0 environment (Javaid et al 2022) characterized by a rapid and unconventional informational flux, new informational platforms and tools and new, challenging online informational consumerism behaviors. Online falsehood has a considerable impact on the behavior and attitudes of the public opinion (Allcott & Gentzkow, 2017). For this reason, fabricated stories, as part of the disinformation phenomena, represents nowadays a global problem which needs to be addressed properly (Pal & Banerjee, 2019).

Even though fabricated stories have always been encountered throughout history (such as Pizzagate) (Edson, et al., 2018, p. 137), their number has grown in recent years as a result of the new era of Internet hoaxes (LaGarde & Hudgins, 2018). Complementary, one of the phenomena that influenced the massive dissemination of fabricated stories in the last years is the so-called phenomena of infodemics, recently defined by the World Health Organization as a “superabundance or excess of information, including false or misleading information, regarding a topic” (World Health Organization, 2022). In this situation, true content is being mixed with false data (World Health Organization, 2022) in order to confuse the user. Infodemic contexts are fueled by the digital communication system (Del Mar Rivas Carmona & Vaquera, 2020) which contributes, especially in crisis situations such as the COVID-19 pandemic, to an accelerated spread of the misleading messages and hoaxes. As a consequence, disinformation, as well as misinformation affects the digital information consumption and generates negative effects on the social behavior of individuals (García-Marín, 2020)

In addition, the young generations (such as Generation Y and Z) get their information primarily online and trust the data available on different social platforms and networks (Shifman, 2013). As a consequence, this new form of information consumption transforms young people into likely victims of Internet hoaxes and online disinformation disseminated on online communities and circles (Perez-Escolarr, et al., 2021, p. 3).

Therefore, as part of the natural process of adapting to the current reality, each person needs to acquire the skills and competencies necessary to conduct self-fact-checking processes, in order to ensure they get informed in a correct and opportune manner (Mantzarlis, 2018). Moreover, the response to disinformation and misinformation must be complex, since these phenomena are not only the problem of mass media outlets, but they are also a social phenomenon, which requires a response to a democratic problem (rather a simple response to a problem of lack of credibility of media institutions) (Persily, 2017). Therefore, the response should be constructed by a matrix of actors, including the main news and media actors, as well as social networks and other technology companies. In this equation, developing such a response requires skills specific to the journalistic specialization, as well as new methodologies and tools to cope with the demands of a continuous developing information system (Herrero & Herrera Damas, 2021, p. 53).

Fact-checking, as an activity, has been historically associated with journalism, reflecting its professionalization in the 20th century, as a "fact-centered" discipline (Graves & Amazeen, 2019, p. 3). However, in time, fact-checking developed into a broader infrastructure consisting of people, organizations, routine and practices, principles and tools which work interconnected to ensure that the general public is better informed (Graves & Amazeen, 2019, p. 1) (Cotter, et al., 2022, p. 2).

Defined as an activity, fact-checking is considered to be a "continuous, consolidated practice of checking the veracity of public discourse" (Herrero & Herrera Damas, 2021, p. 51). Complementarily, in accordance with the author Alexios Mantzarlis, the fact-checking process can be also defined as "a scrupulous analysis driven by one simple question – 'How we do know that?'" (Mantzarlis, 2018, p. 84). Fact-checking cannot, therefore, be considered a simple "spell-checking process, since there does not exist a dictionary-style guidebook comprising all the possible facts or a software solution that can examine all documents and flag anytime something has been misstated as fact" (Mantzarlis, 2018, p. 84). In recent times, the fact-checking does not limit itself only to correctly informing individuals, but extends to monitoring, spotting and disproving any piece of information.

Fact-checking plays at present an essential role in the so-called "post-truth" media landscape (Cotter, et al., 2022, p. 2). In the 21st century, fact-checking begun to revolve around ensuring institutional accountability (Graves & Amazeen, 2019), as a result of a several claims in the political domain that proved to be false and the claims of the Bush administration with regards to the weapons of mass destruction as justification used as a justification for the Iraq War (Marietta, et al., 2015, pp. 578-579). Technological developments gave Internet users the opportunity to develop news-like content, and also allowed independent fact-checking sites (such as Snopes.com, Maldita.es) to establish in order to help dispel conspiracy theories and rumors while also trying to fill the role of watchdogs for politicians, journalists and other public figures (Cotter, et al., 2022, p. 3).

Fact-checking was initially the prerogative of media outlets but it has since extended to nonprofit entities, think tanks, nongovernmental organizations and academic institutions which joined the community of fact-checkers (Stencel, et al., 2022). The increased level of disinformation disseminated during the 2016 U.S. elections demanded further engagement for the fact-checking community, which resulted in a 200% increase in the number of fact-checking entities (Fischer, 2020). In accordance with the conclusions of the Duke Reporters' Lab annual fact-checking census, in 2021 there were 391 active fact-checking projects, 378 of which were still operational in June 2022 (Stencel, et al., 2022).

In 2015 the International Fact-Checking Network (IFCN) was established at Poynter, "at the initiative of the checkers themselves who started to meet informally in 2014 in order to exchange good practices and also errors" (Herrero & Herrera Damas, 2021, p. 66). The main objective of IFCN is to provide a platform which brings together the growing community of fact-checkers worldwide and to advocate for factual information in the global fight against disinformation (Poynter). IFCN promotes the excellence of fact-checking to more than 100 organizations worldwide through advocacy, training and global events (Poynter) and also developed a code of principles as an instrument of accountability to guide fact-checkers worldwide so as to ensure a nonpartisan and transparent verifying process (Perez-Escolarr, et al., 2021, p. 4)[51].

At present, fact-checking is considered an essential instrument for combating the negative effects of false information, especially in the online environment. While past works demonstrated that fact-checking corrections can create a backfire effect, and reinforce the original and inaccurate beliefs of the public opinion (Nyhan, et al., 2013), more recent research in the field showed that fact-checking registered improvements in terms of accuracy of beliefs, supporting the ability to evaluate claims in a correct manner and contributing to the reduction of intentions to share deceiving headlines on social media (Nyhan, et al., 2020) (Yaqub, et al., 2020). Given all the above presented aspects, it can be concluded that fact-checking has the ability to fulfill its core objective of ensuring a correct and well-informed public opinion, representing an essential tool for online platforms, as they have transformed into the main source of information for the general public, keeping abreast of current events and news (Cotter, et al., 2022, p. 3).

Debunking and fact-checking

The literature often employs the terms debunking and fact-checking as almost perfect synonyms, but some distinctions exist.

Definitions:

- **verification** - Some researchers[52] make a difference between fact-checking, which happens once a material is made public and gains public relevance, and verification, which takes place before said material is publicly available. Verification is at the heart of journalistic integrity and focuses not only on the truthfulness of statements, but also on the identity of the producers and transmitters of content, be they human or digital (Balancing Act 2020), so that the sources are transparent and traceable.
- **fact-checking** - Occurs after a material becomes public, and refers to the process of verifying if the facts in a piece of writing or in a speech are correct. To this end, fact-checking employs information from experts, academia, official, governmental institutions (UNESCO Journalism Handbook 2018).

[51] The principles are available at https://ifcncodeofprinciples.poynter.org/know-more.

[52] J*ournalism, 'Fake News' & Disinformation Handbook for Journalism Education and Training*, 2018 (available at **Journalism, fake news & disinformation: handbook for journalism education and training - UNESCO Digital Library);** *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression;* Broadband Commission research report on 'Freedom of Expression and Addressing Disinformation on the Internet, 2020, available at Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression (broadbandcommission.org)

Traditionally, fact-checking is performed by journalists, newsrooms, political analysts. However, at present, due to exponentially increasing volume of public information, fact-checking is also undertaken by independent organisations and NGOs, with a view to holding the authors accountable and informing the public with respect to the validity and factuality of their claims.

- **debunking** - is considered a subset of fact-checking, as it relies on similar skills, but focuses more extensively on fake news and hoaxes (UNESCO Journalism Handbook 2018) and on user-generated content. Pemmet & Lindwall (2021) explain that debunking is not limited to exposing falsehood, but also focuses on instances in which something is presented as less important, less good or less true than it actually is. They state that the overarching objective of debunking is to counteract or minimise the effects of potentially harmful mis- and disinformation" (Pemmet & Lindwall 2021 6; RESIST 2 2021). The objectives of debunking are mainly to: a) assert the truth; b) catalogue evidence of false information; c) expose false information and conspiracies; d) attribute the sources of disinformation; e) build capacity and educate (Pemmet & Lindwall 2021 6).

The most important differences between fact-checking and debunking are reviewed by Pemmet & Lindwall (2021):

a) debunking may be **partisan** (if conducted by governments to expose certain actors), while fact-checking is impartial;
b) debunking is **targeted** on a particular actor or a specific topic. The target is chosen in accordance to the effects it could produce if the mis- or disinformation is left unchallenged; while fact-checking is broad in scope and targets any mis- or disinformation.
c) debunking is **strategic**, as it prioritises its targets and does not focus on everything with equal effort. Some mis- or disinformation attempts, which are not perceived as posing threats to the debunkers interests and/or priorities, are not addressed.

Moreover, debunking also exhibits additional traits:

d) **debate-shaping** as its efforts are directed at preventing or correcting manipulation of public debate.
e) **transparent** regarding the debunker's actions, objectives and funding.
f) **awareness-raising** because they also strive to educate with respect to manipulative techniques (Pemmet & Lindwall 2021 16-17).

Despite the conceptual differences outlined above, more often than not, in practice, fact-checking and debunking work hand in hand. As explained in *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression* (2020 66), fact-checking also employs the proactive debunking techniques in order to expose the process behind the falsehoods, as well as the process through which they have been revealed.

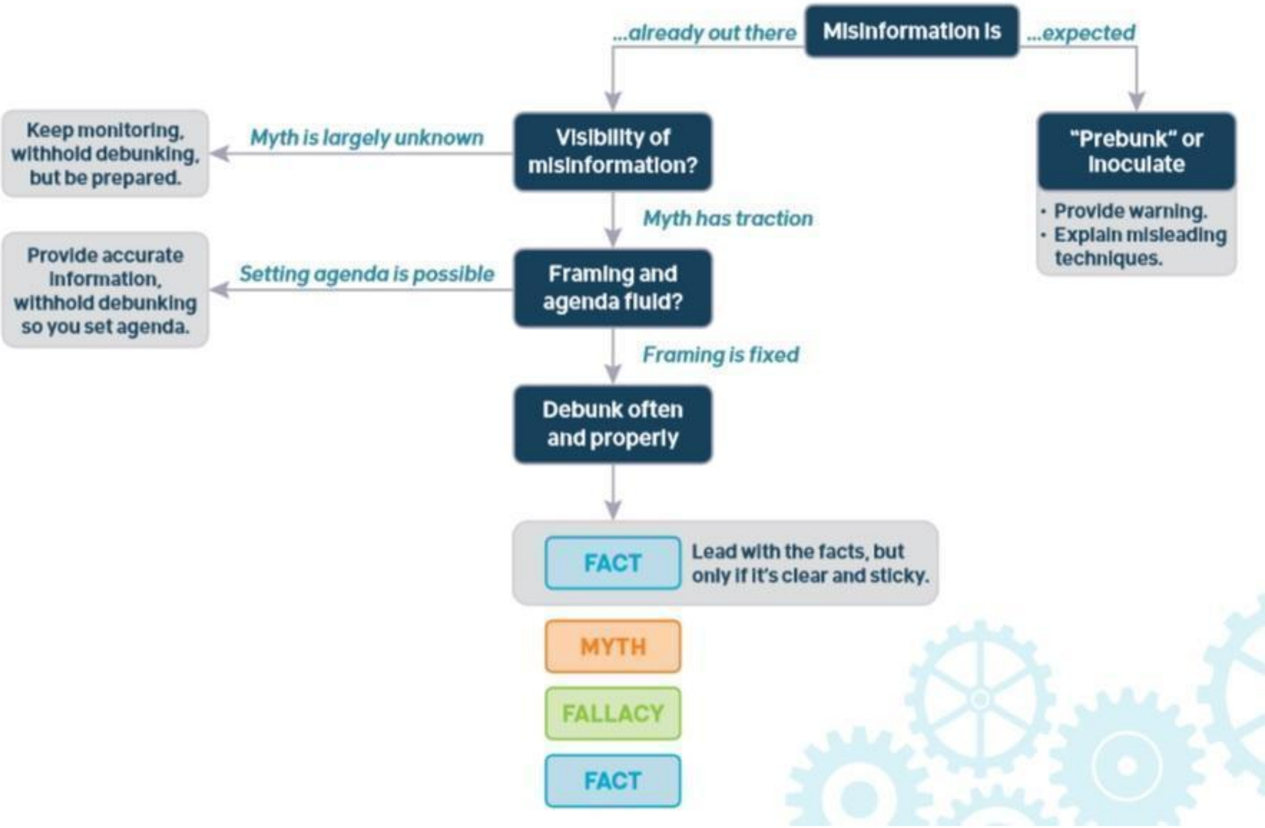
Debunking is guided and informed by the principle that mis- and disinformation should not go unchallenged. To this end, Lewandowsky et al (2020 12) and RESIST 2 (2021 44) recommend using counter-messaging in order to correct, fact-check and debunk. Counter-messaging refers to the construction of corrective messages that maximise clarity and impact and that have a specific step-by-step structure:

1. Fact - the first element in the message is the factual truth;
2. Myth - indication of the false information that is corrected. This false information is only mentioned once.
3. Explain fallacy - it is not enough to say the information is false, an explanation of why it is false, what sort of fallacy is at work is also needed.
4. Fact - the message ends with a restatement of the factual truth because that is what the audience is left with.

In order to minimise the debunking effort, some recommendations are also provided:

1. stick to the subject - do not get distracted into other wider or collateral false narratives, but rather focus on the particular interest in that particular case of mis- or disinformation.
2. use facts, examples and evidence as much as possible and from as many credible sources as there are available.
3. do not engage trolls and their rhetoric because it is counterproductive and time-consuming (RESIST 2 2021 44).

Lewandowsky et al (2020) emphasise the importance of countering disinformation as it appears because, if left unquestioned, it will “stick” due to the emotional appeal it is constructed on and the familiarity backfire effect (i.e., once misinformation is repeated a number of times it is perceived as accurate because it becomes familiar not because of its inherent truth value). Even once disinformation starts spreading, debunking can help curb its dissemination, however, the communicator must choose which pieces of disinformation to tackle because human fact-checking resources are limited and should be used on the most potentially destructive myths.


Misinformation is
...already out there
...expected
Visibility of misinformation?
Myth is largely unknown
Keep monitoring, withhold debunking, but be prepared.
Myth has traction
Framing and agenda fluid?
Setting agenda is possible
Provide accurate information, withhold debunking so you set agenda.
Framing is fixed
Debunk often and properly
"Prebunk" or Inoculate
· Provide warning.
· Explain misleading techniques.
FACT
Lead with the facts, but only if it's clear and sticky.
MYTH
FALLACY
FACT

(Figure 8 The debunking process, source *The Debunking Handbook 2020*, page 8)

Main challenges of debunking and fact-checking

Debunking can be considered a "retroactive" approach, aimed at correcting factually wrong information, when this is already in circulation. There are opinions that claim debunking may have been useful in a time in which there were few written outlets and people would read the same newspaper every day, but that its efficiency has declined considerably with the advent of the new forms of media. Vosoughi, Roy and Aral (2018) have shown, by studying tweets on politically relevant topics, that fake news travels much faster and to a much wider audience than the rectification that follows: it took information that corrected falsehoods six times more to reach the same number of people and twenty times more to be shared a similar number of times. Swire et al (2017) also showed how some people were unwilling to change their support for a candidate to political office in response to being shown correct information about a false statement made by that candidate, if they previously believed that statement was true. Arcos et al. 2021 have identified some serious challenges that need to be considered for advancing in the fight against information manipulations and misleading content. These challenges can be grouped into the following issues:

- The limits of fact-checking practices from the perspective of the audience/publics: This includes factors like acceptance of the verified information once the receiver has been exposed to the misleading content or the mis/disinformation piece.
- Persistence of the falsehood or erroneous/biassed information even after exposition to the fact-checked content or news story.
- Scope and implications of the so-called backfire effect.
- Real impact of fact-checking and debunking: cognitive effects (understanding, retention), but also those associated with attitude change/reinforcement and behavioural (e.g., medium and long-standing behaviours in the consumption of content and access to trusted sources of information and news stories)
- Professional challenges and issues of fact-checking practitioners: the volume of information, speed of dissemination of misinformation stories, perceived political neutrality of the fact-checking organizations, and others.

Looking into more detail at these issues, researchers have noticed that pre-existing attitudes and beliefs play a fundamental role in the acceptance of mis- and disinformation content by audiences (Ewoldsen & Rhodes, 2020). According to the priming theory (Berkowitz, 1984), people react to the messages they received depending on how they interpret the message, the ideas they bring with them and the thoughts that the message evokes. On the other hand, an implication from the cognitive dissonance theory (Festinger, 1962) is that individuals, struggle to accept new information that challenges the previously accepted, and actively seek information that reinforces the previously accepted belief or behaviour –to reduce the dissonance and reinstate the balance

The hostile media phenomenon, or effect, (Vallone et al.,1985) refers to the tendency for partisans to view media coverage of controversial events as unfairly biassed and hostile to the position they advocate" (p. 584). The phenomenon is very relevant for fact-checkers since no matter how much neutral identical journalistic reporting and hard news stories are, the same news coverage of events will be perceived as hostile to their own positions by partisans (Feldman, 2017, p. 2). Walter et al., (2020) argue that holders of partisan positions are more vulnerable to disinformation and misinformation consistent with their own views, but also will likely be more resistant to debunking and fact-checking processes, if the content challenges "pre-existing beliefs, ideology, and knowledge" (Walter et al., 2019).

Several authors (Pennycook et al., 2020; Walter et al., 2020, Nyhan, 2021; Ecker et al., 2022; Ecker et al., 2020) have addressed the so-called backfire effect, which refers to the stronger adhesion to pre-existing beliefs or thoughts when people are confronted with corrective information that challenges these beliefs or thoughts.

Another relevant limitation to the effectiveness of fact-checking is the persistence of the disinformation piece in the memory of the audience and why and how discredited information often continues to influence people's thoughts and behaviours (Lewandowsky et al., 2012; Nyhan & Reifler, 2010). Johnson & Seifert (1994) refer to this persistence as the continued influence effect (CIE) and since then, it has been a topic of attention by researchers (Johnson & Seifert, 1994; Sussman & Wegener, 2022; Ecker et al. 2022; Kan et al., 2021).

With respect to the backfire effect, initial research on the persistence of disinformation found that the backfire effect was widespread. However, subsequent studies suggested that the backfire effect was “extremely rare in practice” (Nyhan, 2021, p.2). Even if this phenomenon is less frequent in scope than initially considered, it should not be disregarded. Confrontation with contrary information from fact-checkers will result in fact-checkers being perceived as hostile to the own positions of holders of partisans views (Feldman, 2017) and can lead partisans to isolate themselves informatively, reject balanced media, and seek ideological similarities, sources, and communities. This in turn, may reinforce their views resulting in greater social polarization.

Addressing these effects is nowadays a significant challenge for fact-checking practitioners, as well as understanding why misinformation and disinformation persist (this being also a critical question for scientists from the communication and psychology fields).

However, more recent studies have demonstrated that the effects of debunking and fact-checking cannot be overlooked entirely. Several recent studies have focused more precisely on the effects that marking information in social media as being inaccurate have on the participants willingness to further disseminate said information to their networks. Chung & Kim (2020) have focused on whether fact-checking can deter the spread of fake news. They relied in their hypothesis on the fact that third person perception (i.e., the perception that while the person themself will not fall prey to fake news, others might) may decrease individuals’ willingness to share that news on social media. The researchers also discovered that fact-checking information also moderates the effects of social media metrics and social sharing intentions, meaning that even if the information appears to have been shared or liked extensively in social media (which could indicate popularity and appeal), if there is a fact-checking warning, individuals will refrain from further sharing it.

Brashier et al (2021) noticed that correcting misinformation may have an effect in the short term, but might fade in the longer term and tested which for of fact-checks produced the longest lasting results. They introduced true or false tags before (prebunking), during (labelling) or after (debunking) the participants read the headlines and concluded that providing the fact-checks immediately after the headlines (debunking) was conducive to longer term retention of the veracity or lack thereof of the information. Their assessment is that debunking is thus correlated with feedback, which boosts long-term retention.

Pennycook & Rand (2017) assessed the role of cognitive reflection in the persistence of misinformation messages. They identified a correlation between the propensity to engage in fake news (negative), on the one hand, or well discern them from real ones and analytical reasoning (positive), on the other hand. In other word, when cognitive reflection exist, people are less likely to engage with fake news. However, the reasoning affects some individuals more than others. Political partisans will be more likely to scrutinize and try to develop counter-arguments against fact-checking contents (Walter et al., 2019). Other authors point out that a rational judgement must be completed with visual information such as images, graphics, and fact-checking labels on social media platforms, and verbal quantifiers of the trustworthiness of the sources (Nyhan, 2021¸ Hameleers et al., 2020; Van der Bles et al., 2020).

Pennycook et al (2020) analysed firstly the implied truth effect which occurs when people who are used to seeing the warnings regarding the veracity of information online, are not presented with such a warning and are therefore more likely to believe the information is accurate, even if, in fact, it is not. This raises serious issues because third parties are needed to examine that information and verify its accuracy, which is an impossible task given the large volume of online available information and creates a dilemma in the public's mind: if the warning is absent, does it mean the information has been verified and deemed accurate or that is has not been verified. The result of Pennycook et al's study indicates that unmarked headlines are viewed as more accurate and are more likely to be shared. Therefore, debunking efforts should be undertaken to provide accuracy warnings regarding trending topics.

Other studies have indicated that polarization on controversial topics can be reduced by explaining the scientific consensus on the topic (Nyhan, 2020). A paradigmatic example that we can find is that of the climate change, where there exists a consensus of 97%; however, only 12% of the American population knows that the consensus is almost absolute, and the critical or denialist scientist community only reaches a 3% (Leiserowitz et al., 2017).

As far as the backfire effect is concerned, it has been discussed by research on radicalization (Day & Kleinmann, 2017) and it has been identified as especially strong within the conservative audience in the USA (Nyhan, Reifler, and Ubel, 2013). Margolin, Hannak, and Weber (2018) highlight the relevant role of the underlying social structure for correcting disinformation and how the existence of a previous relationship between the individual who receives the disinformation corrected and the individual from whom the disinformation comes from, makes the acceptance of the debunked content more likely. The authors did not find differences in the correction of political and non-political rumours. In both cases, it is observed that the corrections from followers and friends are more likely to be accepted.

The backfire effect has also been associated with emotional reactions. Trevors (2022) established a predictive relationship between the refutation of contents and the emotions provoked in individuals. These emotions when they are related to attacks against one's identity result in a negative emotional reaction that anticipates the revision and refutation of the contents. On other hand, negative emotions (e.g., anger or fear) might be particularly likely to evoke the continued influence effect (CIE). Confrontation with information that provokes negative emotions might lead to that individuals experiencing discomfort, and thus that individuals would try to avoid it, by forgetting the (fact-checked) information that is uncomfortable for them. This phenomenon would perpetuate the CIE which is the result of the activation of mechanisms both emotional and cognitive (Susmann & Wegener, 2022). According to these findings, by avoiding potential discomfort through customized fact-checking strategies, fact-checkers may be able to reduce the persistence of misinformation in holders of partisan positions.

Researchers have also noticed that numerous organisations involved in fact-checking and debunking are also developing educational programs to assist the public in developing their abilities to detect mis- and disinformation online. As there is much information to be tackled by small groups of fact-checkers and debunkers, it is important to analyse these educational and popularisation endeavours and the materials they have developed. This will be the scope of section 6. Moreover, it is necessary to develop a strategy of engagement with communities of users as part of the fact-checking endeavour, because this will increase the chances of exposure by individuals to verified content and news stories, not directly from fact-checkers but from friends and relatives.

The profile of a fact-checker

Defining the core skills of the fact-checkers has been the objective of both employers and educators, most of the studies that focus on this topic were based on the hypothesis that journalistic skills and competencies should not be transferred to skill practice (Ornebring & Mellado, 2016). Since the scarcity of skills of the lack of time are two main factors that influence the ability of journalists to verify a certain piece of information and conduct fact-checking activities, it is true to say that skill practice is influenced by time (opportunity) and context (Himma-Kadakas & Ojamets, 2022, p. 870).

The core skills of journalists are usually transferable, with journalists using them in various stages of the news reporting process. However, before focusing on the main skills a fact-checker should possess, it is important to identify and define the fact-checking skills that can be developed during/within a study program, given the fact that nowadays each journalism student learns how to use available journalistic tools and methodologies in order to avoid becoming a victim of the disinformation phenomenon. Therefore, a study conducted by a research team from the Loyola Andalucia University on first year high school students enrolled in a specific course at two universities from Spain, extracted a series of fact-checking skills within the context of social competencies, as briefly described in Figure 2 (Perez-Escolarr, et al., 2021, p. 7).

	SOCIAL COMPETENCIES				
	Ability to use information and communication technologies to communicating, accessing information sources, archiving data and documents to create content, presentating tasks, learning, research and cooperative work	**Ability to integrate knowledge and cope with the complexity of formulating judgements based on information that, being incomplete or limited, includes reflections and decision-making based on evidence and arguments related to the application of**	**Ability to think and act according to universal principles that are based on the value of the person, the cultural heritage and are aimed at the full personal, social and professional development of the student**	**Ability to question things and researching the foundations on which ideas, values, actions and judgements are based and to promote the capacity for initiative in analysis, planning, organization and management**	**Ability to present knowledge in all areas of knowledge, in a clear and unambiguous way, showing interest in interacting with others and ability to maintain a critical and constructive dialogue, as well as to speak in public if necessary**

		their knowledge and judgements			
FACT-CHECKING SKILLS	1. Use and mastery of technological means	11. Cognitive reflection	23. Thought and critical reasoning both deductive and inductive	30. Initiative capacity in analysis, planning, organization and management	38. Retention and synthesis capacity
	2. Expression in the media without grammatical or spelling errors	12. Assertiveness and empathy	24. Participation and social gathering	31. Creativity	39. Teamwork
	3. Analysis of information sources	13. Have an ethical and responsible attitude of respect for people and the environment, with responsible consumption	25. Being able to integrate and work efficiently in multidisciplinary teams assuming different roles and responsibilities	32. Motivation for achievement	40. Social responsibility
	4. Ability to generate audiovisual content	14. Interpretation, argumentation and problem solving	26. Flexibility and/or adaptability	33. Initiative and leadership	41. Dialogue critically and constructively
	5. Management of computer tools	15. Decision-making capacity	27. Integrity and value of the performance of the professional activity	34. Fostering the imagination	42. Ability to interpret, argue and solve problems
	6. Strengthening of research knowledge and skills	16. Social responsibility	28. Tolerance to stress	35. Development of innovative capacity	43. Self-confidence
	7. Capacity for collaboration, cooperation and connectivity	17. Respect the fundamental rights and equality between men and women	29. Self-employment	36. Independent learning	
	8. Manipulative skills and	18. Increased selective attention		37. Identify, practice and project	

	simultaneous tasks	and mental alertness		proactive competition	
	9. Information search and analysis skills	19. Do not incite hatred, racism, homophobia, etc			
	10. Improvement and development of receptive communication.	20. Rejection in the face of hoaxes			
		21. Ensure the veracity of the data			
		22. Information contrast			

Table 7. Fact-checking skills and social competencies (Perez-Escolarr, et al., 2021, p. 7).

In addition to these skills, the studies on journalistic skills and competencies identified the following skills specific for the fact-checker profession:

1. **critical thinking** – necessary for the selection and evaluation of the sources and the verification of information (Carpenter, 2009). This skill also enables professionals to comprehend more in-depth knowledge in a specific subject (Himma-Kadakas & Ojamets, 2022, p. 870);
2. **ability to evaluate newsworthiness** – even though this skill might seem related to critical thinking, it is considered vital in the situations when journalists are exposed to information disorder (Carpenter, 2009). A significant volume of the data that reaches the online environment may appear newsworthy, but a critical evaluation of its factual basis can determine the contrary (Himma-Kadakas & Ojamets, 2022, p. 870);
3. **knowledge of topics outside the journalistic domain** – it can certainly help professionals to recognize potentially false information (Himma-Kadakas & Ojamets, 2022, p. 870);
4. **advanced knowledge in information gathering and investigation** – since verification of information is a crucial step in the process of fact-checking, developed knowledge of tools, instruments and methods to evaluate the source of information and identify the origin of it are essential (Himma-Kadakas & Ojamets, 2022, p. 871);
5. **knowledge of social media** – social media has become nowadays the an essential environment for the derivation and dissemination of valuable and opportune information (Carpenter, 2009). Therefore, a fact-checker should definitely know which are the principles and main functions of social media as a platform (Himma-Kadakas & Ojamets, 2022, p. 871).

All in all, one can say that the profile of the fact-checker differs from the traditional journalistic role, going beyond the traditional journalistic practice. However, this topic has been poorly addressed until now and still requires further research in order to be able to develop a general fact-checker profile that can be applied to all the professionals that fulfill this role in different domains of activity.

Pre-bunking

Given the informational overload to which a person is subjected, studies have found that another efficient strategy to stop the spread of disinformation is a proactive approach, called "pre-bunking" (preemptive debunking). This relies on the idea of "inoculating" people against disinformation, so that they are better trained to identify disinformation tactics when they are faced with them. Those who support pre-bunking (Lewandovsky and van den Linden 2021) believe that, just as in the case of a real vaccine, once a person comes in contact with a "weakened" version of the practice of disinformation, then they will become immune when encountering that practice in the real world.

One of the first mentions of the term pre-bunking in the academic literature on disinformation is the article by Cook (2016), who strongly criticises the debunking approach. He argues that debunking is inefficient because people build mental models in which the false information fits neatly. Once a retraction is circulated, that particular mental model would be incomplete if the new information was accepted and the old one corrected. However, according to Cook, people prefer complete, even if incorrect, mental models over incomplete ones. The alternative is to help people build correct mental models through inoculation, especially by preemptively exposing the logical fallacy employed to spread a particular piece of disinformation.

The PROVE framework (Blastland et al 2020) for evidence-based communication stipulates the five important rules that can help develop informative messaging of a variety of topics, that can keep the audiences engaged and can help educate them even on complex scientific matters. The goal of evidence-based communication is not to persuade, but to inform, and eventually re-empower the individuals.

1. Pre-bunk - using media monitoring and risk assessment (see below) the communicators can anticipate the topics that might lead to mis- or disinformation and make the necessary preparations to preemptively warn the public;
2. Reliably inform - openly, honestly and transparently offering information leads to trust-building in institutions/organisations/companies/governments;
3. Offer balance - evidence must not be presented in a biassed manner or omitted so as to serve somebody’s interests, but as objectively as possible;
4. Verify quality - the evidence must be checked for quality, multiple sources, if available, should be used so as to foster credibility;
5. Explain uncertainty - if the evidence is ambiguous, uncertain, incomplete, these aspects should be openly disclosed.

These steps would ensure that the public feels confident trusting the communicators when they try to warn them with respect to potentially harmful mis- or disinformation and listen to their recommendations more openly than if they tried to use persuasive techniques to reach them. Moreover, Blastland et al (2020) also explain that prebunking can play a decisive role in inoculating people against disinformation but it must be done strategically, that is public forums and popular news sources must be scanned constantly for indicators of what the public is concerned with so that the topics for disinformation might be detected ahead of time, before they become engrossed in disinformation. Thus strategies for evidence-based communication on those topics could be devised in a timely manner.

Before the advent of the online games (see section 5.1), the inoculation approach was conducted through in-class teaching of the controversy on various topics or through Massive Open Online Courses. Students taking part in classes involving pre-bunking strategies were then able to identify the main myths in their assignments.

A series of studies summarised by Lewandowsky and van den Linden (2021) have shown the efficiency of inoculation against fake news. Thus, van den Linden et al. (2017) and Cook et al. (2017) both conducted an inoculation experiment where people were presented with disinformation about climate change as well as an inoculation treatment through warnings about disinformation techniques. Those who had received "inoculation" before seeing the particular piece of disinformation tended to rate the accuracy of the false statements as much lower than those that had not been exposed to the inoculation treatment.

Pre-bunking also exhibits certain limitations. Firstly, it is moderated by partisanship as the effect is diminished for people from one side of the political spectrum. The second limitation concerns the setting: pre-bunking interventions are considerably more efficient in a laboratory setting than in the real world. Finally, the inoculation tends to wane after a while, as people are again exposed to the usual disinformation (Roozenbeck and van den Linden 2022).

Inoculation or pre-bunking has shown consistent results in stopping people from believing and sharing disinformation. By giving people a forewarning about the strategies that actors spreading disinformation use, pre-bunking convinces people to stop and think about what they are seeing, attempt to rate the accuracy of a piece of news and evaluate whether this is worth sharing further.

Reverse psychology

Reverse psychology has led to the development of another strategy aimed at creating awareness and resilience to propaganda and disinformation.

As a persuasion tactic, reverse psychology has been extensively used in marketing and is based on encouraging the target audience to do what is desired by advocating the opposite behaviour in a way that makes the alternative more convincing and alluring. Illustrative examples are offered by the study of social influence tactics under the compliance paradigm and include:

- foot in the door, in which compliance with a small request increases compliance with a later, larger request
- door-in-the-face, in which noncompliance with a relatively large first request increases compliance with an immediate, smaller request;
- and disrupt-then-reframe, in which a request is phrased in unconventional terms then reframed to the advantage of the influence source. (Donald, Nail, Harper, 2010, 1)

Such tactics are normally used to favour influence in unbalanced power distance, in which the less powerful actor may wish to introduce conformity to its suggestions. And while acting against self-conformity to true intentions may be often counterintuitive, the success of such tactics, as evidenced by field literature on the model of social response, prove that it is noteworthy documenting strategies of persuasion that may anticipate the behaviour of the other actor engaged in communication. (Nail & Van Leeuwen, 1993; Willis, 1965).

Serious gaming, a concept that shall be studied extensively in another chapter, has allowed pairing of the reverse psychological tactics with the therapeutic paradox, namely the use of the therapist of a situation "that can only be controlled by using direct communication and by abandoning indirect tactics" (Klein, 1974), that is by renouncing pathological behaviour patterns.

It has been proven that we can use the development of a psychological paradox situation to apply reverse thinking and expose target audience to a deeper understanding of the manipulative tactics used by propaganda and disinformation outlets. Think of a situation in which one takes his understanding of news from alternative channels (clones of news portals, grey zone sites and blogs, promoted via bots etc.). When exposed to a situation in which one has to step in the shoes of the propaganda and/or disinformation agent and use disinformation tactics, the first expected outcome is to grow resilience to tactics to which he/she has familiarised via debate, study, practice and serious gaming. In psychological terms, this equals to creating a situation that cannot be mitigated via old patterns of information collection and in which the individual needs to react in full awareness and defence to his/her exposure to false information and manipulation techniques. Hence, the expected cognitive and behavioural change. However, in order to favour behavioural change through reverse psychology techniques employed in e.g. serious gaming tactics, one must "incorporate sound cognitive, learning, and pedagogical principles into their design and structure" (Greitzer, Kuchar, Huston, 2007, 1).

Main challenges and means to overcome them

In an age where social media has become the main source of information for almost all social categories, the fact-checking process represents nowadays an essential step to be followed in order to ensure a correct judgment of the credibility of information obtained from the Internet. Given the technological boom which characterized the last decade, that facilitated the development of multiple technological instruments and tools to be used online for a both positive and negative output, the increased level of fake news sites, hoaxes and misinformation online is now considered a concern and, to some extent, a security issue (Stenger, 2016).

In this context, being able to make the difference between reality and manipulated information represents no longer a capability self-developed during lifetime, but is nowadays defined as a necessary skill to be trained and practiced with regularity in order to keep up with the evolution of the volatility of the online environment. As a consequence, the fact-checker has become a key profession in both informative/communicative and democratic processes that occur in contemporary society (Herrero & Herrera Damas, 2021, p. 51), with professional fact-checkers representing an essential factor for the control of information that is disseminated/circulated in the online environment (Herrero & Herrera Damas, 2021, p. 49).

Some experiments have been conducted to evaluate the effects of the exposure of disinformation on target audiences. These effects cover three different sets of effects: cognitive, emotional and behavioural.

One of the main conclusions of these studies, points out that, although the effects of disinformation depend on several factors, pre-existing attitudes and beliefs play a fundamental role on the acceptance of malicious content, such as disinformation narratives, by individuals (Ewoldsen & Rhodes, 2020). However, unless pre-exposure and post-exposure surveys are conducted with some frequency, it will be difficult to assess real impact and effective influence on attitudes and behaviour.

According to Arcos (2018), more evaluative research based on social research techniques is necessary to provide findings “on the cognitive/informational impact (message exposure, understanding, and retention), attitude impacts (attitude creation, modification, and reinforcement), and behavioural effects (how people will behave or will cease to behave as a consequence of accepting those malicious messages). It is of utmost importance to track these evidence-based assessments to evaluate the medium and long-term effects of mis- and disinformation in societies. Valkenburg & Oliver (2020) discussed the need for assessing reliable data on audiences and their exposition to disinformation messages –noticing the information can come from multiple devices or channels– for a full comprehensive understanding of its impact. In the same way, more research on the impact that the fact-checked content has on the publics is needed. Although some evaluation research has been conducted, it has been focused mainly on electoral processes (Wintersieck, 2017). Measurement of the web-traffic for the fact-checking organizations and tracking of information flows of fact-checked content in social media can be useful, but do not cover the full tracing of fact-checked contents.

Lukito’s research (2020) on the activities of Internet Research Agency (IRA), and how these were coordinated through the different social media, suggests the interest of conducting post-mortem analysis (Arcos 2018) covering full tracking (platform by platform) in order to detect what channels and how they have been used, and develop ad-hoc fact-checking strategies.

The growth of disinformation and misinformation and its expansion to non-political issues has relaxed the debate on whether fact-checking activities should be carried out by newspapers or independent organizations. Fact-checking was always part of the journalistic process but the new information environment has raised the need for verification on a number of issues that surpass traditional political news stories. The rise of new fact-checking organizations, as a result of the higher demand for verification related to the COVID-19 associated infodemic, seems to have affirmed the principle of political neutrality; a high number of fact-checking organizations are NGO’s and hence this principle is not questioned in the same ways as newspapers.

Unlike traditional news outlets that also have sections on opinion, and develop news coverage on specific events and development, setting the agenda on political and international news the fact-checkers emerge as an active listening agent that identifies and satisfies the information needs of its community. Fact-checking organizations are the result of the new paradigm of journalism marked by technological development and citizen interaction in information activity, a model that, as Pisani (2008) points out, breaks with the hierarchical communication model, "from one to many", moving to a horizontal communication model and "from many to many".

From this approach, a mutually beneficial relationship is established between the organization and its consumers. Involving audiences in the different phases of fact-checking can be essential to fill the existing gaps in organizations by expanding the topics on misinformation and disinformation, identifying viral information that is disseminated in closed messaging platforms, and contributing to the dissemination of corrected information. The commitment to programs such as *Superpoderes* of the Spanish foundation Maldita is a good example of how the knowledge of audiences (on health, politics, climate, International Relations, history, technology, etc.) can be integrated into organizations supporting the phase of consulting independent experts to interpret the data (Graves, 2017).

A recent challenge that these organizations have to deal with is the appearance of informal or non-professional fact-checkers that from twitter mainly disseminate misinformation or disinformation corrected, a phenomenon that has grown up with the war in Ukraine. While in some cases, these profiles can help professional fact-checkers, it is difficult to determine possible cover intentions. Barriers to entry must be established such as the monitoring of methodologies, expert knowledge contrasted in a certain scope, verification of professional credentials and reputation.

On other hand, the growth of disinformation activity in recent years has been linked to a boom in research and development of tools aimed at automating certain tasks that are necessary for the verification process. This automatic fact-checking is based on the development of algorithmic models based on deep learning, machine learning, natural language processing (NLP) and big data (Huynh & Papotti, 2019, Miranda et.al, 2019, García-Marín, 2022). These technologies support the detection of factual claims worth verifying, check if a content has already been verified and perform affirmation validation to determine the feasibility of the detected content (Miranda et al., 2019). Communication with audiences can benefit from intelligent chatbots (Cha et.al, 2020) that incorporate adversarial generative networks (ADNs) to retrieve and generate evidence and explanations in natural language. These tools allow the user to check if a content has already been verified as well as send content suspected of misinformation or disinformation to be verified.

The attention for the development of models to support verification has expanded in recent years, as a result of the boom in demand for content verification because of the Covid-19 pandemic, however there are multiple challenges still pending. For example, García-Marín (2022) points out the absence of models for analysis of fake audio, as well as a poor attention to the detection of fake images and video. Likewise, the changes in the diffusion patterns and the entry into this scenario of social networks such as Tik Tok or Telegram, newer or scarcely used for the dissemination of these contents, means that the tools for its exploitation are still in an emerging phase.

Inspiring practices

This section will briefly present hands-on approaches, handbooks and toolkits developed by specialists in the field of communication, who have been confronted with disinformation in their activities and have developed means of combating it, which could prove valuable resources for communicators engaged in countering disinformation.

RESIST 2: Counter-disinformation toolkit overview (available at RESIST 2 Counter Disinformation Toolkit - GCS (civilservice.gov.uk)) an freely available toolkit for communicators to help them develop the necessary skills and competences to tackle disinformation and its effects on companies, campaigns, society as a whole.

6 stages each with hands-on instruments to operationalise them into actionable steps:

1. **Recognise** - understand the types of disinformation that exist in the overcrowded media environment at present and the potential dangers and threats they pose. In order to recognise disinformation effectively, one needs first to analyse the messages produced. Secondly, the narratives behind the messages need to be identified, as well as the values, identities, beliefs that they reflect (the writers of RESIST 2 call this the brand). Once the brand is identified the interests can also be ascertained, as well as the potential impact of disinformation on target audiences.

The FIRST indicators for analysing the message:

a) Fabrication - Is there any manipulated content? E.g., a forged document, manipulated image, or deliberately twisted citation.
b) Identity - Does anything point to a disguised or misleading source, or false claims about someone else's identity? E.g., a fake social media account, claiming that a person or organisation is something they are not, or behaviour that doesn't match the way the account presents itself.
c) Rhetoric - Is there use of an aggravating tone or false arguments? E.g., trolling, whataboutism, strawman, social proof, and ad hominem argumentation.
d) Symbolism - Are data, issues or events exploited to achieve an unrelated communicative goal? E.g. historical examples taken out of context, unconnected facts used to justify conspiracy theories, misuse of statistics, or conclusions that are far removed from what data reasonably supports.
e) Technology - Do the communicative techniques exploit technology in order to trick or mislead? E.g. off-platform coordination, bots amplifying messages, or machine-generated text, audio and visual content. (RESIST 2 2021 10-11)

2. **Early warning** - overview of the tools available to spot disinformation in a timely manner and monitor the media environment

The first element that ensures early detection is monitoring the risks, which can be done using:

a) Platform analytics Each social media platform has an analytics function that provides data on accounts or pages that you own. Platforms that you own pages on are an important source of insight for understanding how people engage with your content.
b) Google Trends Shows how frequently terms are searched for on Google. The results can be broken down by time, country, and related queries to focus attention on a specific timeframe, location, and/or topic. This is useful for revealing spikes in interest and can help

guide your attention to specific days, locations or topics where interest in a debate has changed.

c) TweetDeck Create a Twitter dashboard to follow multiple timelines, accounts and search terms in real time. Note that you can monitor accounts and keywords in Tweetdeck without being a follower. Available at tweetdeck.twitter.com.

d) Browser extensions There are a number of apps that can be added to your browser to speed up or even automate functions such as translation, image searches and taking screenshots. This is especially useful for speeding up simple tasks that you need to do often. (17)

The second element is to develop contingency plans in case disinformation affects priorities such as: objectives, information, brands, audiences and come up with detailed scenarios as to how these could be affected and how the communicators could respond to those threats in an effective manner.

3. **Situational insight** - refers to the ways in which communicators can turn information into actionable insight for decision-makers. More precisely, the information resulting from stage 2, needs to be presented to decision-makers in such a way that it becomes relevant and can guide and inform decisions.

4. **Impact analysis** - presents the structural analysis techniques that can assist communicators in predicting the potential impact of disinformation and produce objective assessments.

When determining the impact, several key aspects must be taken into account and measured in order to obtain an informed and objective assessment rather than a gut feeling.

a) What is the degree of confidence? the results from the monitoring stage should be treated as indicators of possible trends rather than fixed and determined opinions and should be evaluated in terms of risk (high, medium, low) and likelihood (high, medium, low).

b) How does mis- or disinformation affect your areas of responsibility? Clearly articulating potential consequences assists in identifying the best responses and resilience building methods.

c) How does the mis- or disinformation affect your communication with the public? Communicators need to use the FIRST indicators to answer this question.

d) How does the mis- or disinformation affect your brand? what interests/values/beliefs might be targeted, why and how?

e) What is the likely reach of the mis- or disinformation? How far can it spread and what kind of audiences can it reach?

f) How should I prioritise the mis- and disinformation? There is no need to address every single piece of disinformation that appears, since it is also impossible from a resource point of view. If the monitoring stage and the impact analysis are done appropriately, then it is easier to identify the disinformation that has potential to impact the brand in the most significant manner and needs to be addressed. “A prioritised response is one in which there is a clear and compelling need to protect government objectives, information, brands and/or audiences.” (29)

5. **Strategic communications** - maps the communication skills that could be employed to develop communication strategies meant to increase credibility, create proactive, engaging content for the target audience. If communication to correct disinformation is needed, then certain rules apply.

a) Follow communication best practice (*OECD Principles of Good Practice for Public Communication Responses to Mis- and Disinformation* recommended): transparency, inclusiveness, responsiveness, whole-of-society, public-interest driven; institutionalisation; evidence based; timeliness; prevention; future-proof.

b) What are my communication options? Communication could take place on traditional channels (radio, television, print newspapers) or on digital platforms or on social media. Moreover, the options also include proactive versus reactive methods. Among the proactive efforts are mentioned: inoculation, awareness raising, campaigns, network building, counter-brand, resilience building (RESIST 2 2021 40). Among the reactive efforts are enumerated: debunking, counter-narrative, crisis communication, policy response (RESIST 2 2021 43).

6. **Tracking effectiveness** - tools to measure the effectiveness of strategic communication campaigns. The measurement should take place along two different coordinates: outputs and outcomes. Outputs refer to the messages created and disseminated and will be measured with respect to audiences reached and engaged. Outcomes refer to the impact of that communication on the world and will be measured by tracking changes in the target audience's behaviours and thinking.

MSB Countering information influence activities A handbook for communicators. Swedish Civil Contingencies Agency (MSB), 2019 (available at Countering information influence activities : A handbook for communicators (msb.se))

The authors of the handbook define information influence activities as involving “potentially harmful forms of communication orchestrated by foreign state actors or their representatives. They constitute deliberate interference in a country’s internal affairs to create a climate of distrust between a state and its citizens” and through the use of deception to undermine democracy.(11-12) The handbook provides very useful guidelines for communicators from company/organisation/institution/government to employ in order to counter such activities. Disinformation is considered to be a type of information influence activities and therefore the handbook guidelines can be used to counter it as well.

In order for a communicator to counter information influence activities and campaigns, they must follow certain guidelines:

PREPARE The preparation stage is very important in order to assess the risks and threats that the company/organisation/institution might be or become subject to so that mechanisms are developed to ensure an appropriate, adequate and fast response. The preparation stage focuses on three aspects:

a) **Raising awareness** - if the public in or outside the company/organisation/institution/society as a whole is aware of the issues, of the possible threats and vulnerabilities, then they are more willing to work together, pool their resources and knowledge and create a more comprehensive resilience-building approach.

b) **Building trust** - Influence operations and disinformation campaigns are aimed at subverting trust and disengaging audiences. Therefore, strategic communications’ main role is to promote the company/organisation/ institution/government values, vision, objectives, etc., through consistent, proactive, positive, correct, well devised and distributed messages, employing well-defined narratives that promote the core values.

c) **Assessing risks and vulnerabilities** in order to know where the company/organisation/institution/society as a whole may be most exposed to disinformation attacks: “ with a specific focus upon vulnerable stakeholders/audiences, key values, messages and narratives, and the overall risk to your organisation’s core activities” (34)

ACT Responses need to be adapted to the type of organisation, to the public, to the channels they are transmitted through, to the type of influence activity that targets the organisation.

a) **Choose your response**. The authors of the handbook propose four possible types of responses from which to choose: assess, inform, advocate, defend. The first two responses fit into a broader category of fact-based responses, while the latter two can be framed as advocacy-based responses.

   Fact-based responses have two levels. Level 1 is Assess and it focuses on mapping the situation, fact-checking and investigating transparently, while Level 2 is Inform and centres around making a statement, correcting, referring independent sources that can corroborate the information, asserting values, notifying stakeholders, issuing a holding statement (MSB Handbook for communicators, 2020 36).

Advocacy-based responses also have two more levels. Level 3 Advocate which includes dialogue, facilitation, multipliers (engaging with key communicators to disseminate further), piggybacking (using existing events, initiatives, or debates to promote the facts of the case), formal statements and storytelling. Level 4 Defend comprises of ignoring (i.e., doing nothing about the disinformation), reporting (to the appropriate authorities), blocking (the user who promoted the disinformation), exposing (the actor behind the disinformation) (MSB Handbook for communicators, 2020 37).

b) **Check your facts.** When countering disinformation and influence operations, facts matter because they support the legitimacy of the company/organisation/ institution/government, which cannot be seen as engaging in the same type of deception. Therefore, regardless of the type of response adopted, facts should form and inform the basis for the narratives and messages.

c) **Use social media**, with its built-in functions of tagging, notifications, links and attachments both to be aware of the messages circulating regarding the company/organisation/ institution/government, but also to build networks of transmission, which could be activated and engaged when a disinformation campaign needs to be shut down.

LEARN Communicators need not only prepare and act, but also evaluate the measures taken and assess their efficacy.

a) **Describe**. Actively describing the situation in which the company/organisation/ institution/government was targeted by disinformation will help to establish an organisational understanding of the event, a continuity for best practices, and to design other possible proactive measures.

b) **Reflect.** Analysing the effects of the disinformation, of the responses, weighing the good and the bad outputs and outcomes could also increase preparedness for future events

c) **Share.** Experiences and expertise need to be shared among colleagues and with management so that there is a common understanding and culture regarding mitigating risks and vulnerabilities posed by disinformation.

First Draft[53] is one of the first global initiatives to focus on “providing practical and ethical guidance in how to find, verify, and publish content sourced from the social web.” The initial founding partners are: BellingCat, Dig Deeper, Emergent.info, EyeWitness Media Hub, Google News Initiative, Meedan, Reported.ly, Storyful, and VerificationJunkie) and in 2016 it expanded to become an international Partner Network of newsrooms, universities, platforms and civil society organisations.

[53] About (firstdraftnews.org)

The International Fact-Checking Network[54] (IFCN) at Poynter was also established in 2015 and it is a global leader in fact-checking, promoting best practices and fact-checking standards included in the IFCN's Code of Principles (see chapter ____).

MediaWise[55] is also developed by Poynter and focuses on enhancing young generations' and not only abilities to detect disinformation in online content, such as critical thinking and media literacy and become critical consumers of online content. Their motto is: We believe that when facts prevail, democracy wins.

Debunk EU[56] is an independent technological analytical centre and an NGO. They research disinformation in the public sphere and conduct media literacy educational campaigns. They employ artificial intelligence to research disinformation and bridge the gap between the speed with which disinformation is produced and debunked.

The European Digital Media Observatory[57] (EDMO) is a hub for fact-checkers, academics and other relevant stakeholders which allows them to work together and actively links media organisations, media literacy experts, policy makers, teachers and citizens so that actions taken to fight against disinformation are coordinated.

NATO Stratcom Center of Excellence[58] is multi-nationally constituted and NATO-accredited international military organisation, which is not part of the NATO Command Structure, nor subordinate to any other NATO entity. Its mission is to contribute to the strategic communications capabilities of NATO, NATO allies and NATO partners. It is made up of multinational and cross-sector participants from the civilian and military, private and academic sectors, trainers, educators, analysts and researchers.

UN Verified Campaign[59] appeared in the context of the COVID 19 infodemic and it was aimed at flooding the online space with facts regarding the pandemic so as to counter the numerous disinformation campaigns. They also provide guidelines for creating evidence-based communication campaigns tailored to penetrate the saturated social media environment.

Anti-Fake[60], Romania is an initiative to promote digital education, raise public awareness and counter disinformation. It produces a series of products and analyses in order to enhance the general public's abilities to identify disinformation and its manifestations.

Misreport[61], Romania is a weekly report on the latest attempts at disinformation and their debunking written by two prominent Romanian journalists.

[54] About IFCN - Poynter

[55] MediaWise - Poynter

[56] About Debunk EU | Debunk

[57] European Digital Media Observatory (EDMO) | Shaping Europe's digital future (europa.eu)

[58] StratCom | NATO Strategic Communications Centre of Excellence Riga, Latvia (stratcomcoe.org)

[59] Verified | Home (shareverified.com)

[60] Despre proiectul antifake.ro | antifake.ro

[61] Misreport | Revue

Factual[62], Romania was started in 2014 and is the first fact-checking site for public policies and statements in Romania. It is entirely funded by readers' donations and does not accept any form of government funding.

InfoRadar[63] is a news platform as well as an educational instrument for citizens, initiated by the Romanian Ministry of Defence. It is updated by a community of military public relations officers and journalists and its goal is to become a new public communication channel to assist in correctly informaning the public on topics pertinent to the Romanian Armed Forces and to correct non-factual debates in the public arena with respect to the Romanian Armed Forces.

References:

1. Allcott, H. & Gentzkow, M., (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, p. 211–236
2. Arcos, R. (2018). "Post-event analysis of the hybrid threat security environment: assessment of influence communication operations." *Hybrid CoE Strategic Analysis* 12. https://www.hybridcoe.fi/wp-content/uploads/2020/07/Strategic-Analysis-2018-12-Arcos.pdf
3. Arcos R, Gertrudix M, Arribas C and Cardarilli M. (2021). "Responses to digital disinformation as part of hybrid threats: a systematic review on the effects of disinformation and the effectiveness of fact-checking/debunking" *Open Research Europe*, 2(8). https://doi.org/10.12688/openreseurope.14088.1
4. Basol, Melisa, et al. (2021) "Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation." *Big Data & Society* 8 (1): 20539517211013868.
5. Basol, Melisa, Jon Roozenbeek, and Sander Van der Linden (2020).. "Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news." *Journal of cognition* 3 (1)
6. Berkowitz L. (1984). Some effects of thoughts on anti- and prosocial influences of media events: A cognitive-neoassociation analysis. Psychol Bull. 95(3): 410–427.
7. Blastland, M., Freeman, A. L., van der Linden, S., Marteau, T. M., & Spiegelhalter, D. (2020). Five rules for evidence communication.
8. Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5), e2020043118.
9. Carpenter, S., (2009). An Application of the Theory of Expertise: Teaching Broad and Skill Knowledge Areas to Prepare Journalists for Change. *Journalism & Mass Communication Educator*, 64(3), p. 287–304.
10. Cazalens, S. et al., (2018). *A Content Management Perspective on Fact-Checking.* Lyon, France, s.n., pp. 565-574.
11. Chan, M.-p. S; Jones, C.R.; Jamieson, K.H; Albarracin, D. (2017). "Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation". *Psychol. Sci.* 28:1531–1546
12. Chung, M., & Kim, N. (2021). When I learn the news is false: How fact-checking information stems the spread of fake news via third-person perception. *Human Communication Research*, *47*(1), 1-24.
13. Ciampaglia, G. L. et al., (2015). Computational Fact Checking from Knowledge Networks. *Plos One*, pp. 1-13.
14. Compton, Josh, et al (2021):. "Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories." *Social and Personality Psychology Compass* 15 (6): 1-16
15. Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence". *PLOS ONE*, 12(5), e0175799. https://doi.org/10.1371/journal. pone.0175799

[62] Factual.ro - Factual • Adevărul din politică

[63] Inforadar - Portalul stirilor corecte :: MApN

16. Cook, John, et al (2022). "The cranky uncle game—Combining humor and gamification to build student resilience against climate misinformation." *Environmental Education Research*: 1-17 (online first)
17. Cook, John (2016). "Countering climate science denial and communicating scientific consensus." *Oxford research encyclopedia of climate science*., https://oxfordre.com/climatescience/abstract/10.1093/acrefore/9780190228620.001.0001/acrefore-9780190228620-e-314, accessed 5.09.2022
18. Cotter, K., DeCook, J. R. & Kanthawala, S., (2022). Fact-checking the Crisis: COID-19, Infodemics, and the Platformization of Thruth. *Social Media + Society*, pp. 1-13.
19. Day, J. & Kleinmann, S. (2017). "Combating the Cult of ISIS: A Social Approach to Countering Violent Extremism."*The Review of Faith & International Affairs*, 15(3): 14-23, https://www.doi.org/10.1080/15570274.2017.1354458
20. Del Mar Rivas Carmona, M. & Vaquera, M. L. C., (2020). Pandemic and post-truth: The impact of COVID-19 on Whatsapp communication. *Prisma Social* , pp. 110-154.
21. Dobbs, M., (2017). *The Rise of Political Fact-checking: How Reagan Inspired a Journalistic Movement: A Reporter's Eye View,* s.l.: New America Foundation .
22. Ecker, U.K.H., Lewandowsky, S. & Chadwick, M. (2020). "Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect." *Cognitive Research* 5(41). https://doi.org/10.1186/s41235-020-00241-6.
23. Ecker, U.K.H; Lewandowsky, S.; Cook, J.; Schmid, P., Fazio, L.Z.; Brashier, N.M.; Kendeou, P.; Vraga, E. K. and Amazeen, M. A. (2022). "The psychological drivers of misinformation belief and its resistance to correction." *Nature Reviews Psychology*, 1. https://doi.10.1038/s44159-021-00006-y
24. Edson, C. T. J., Zheng, W. L. & Richard, L., (2018). Defining "Fake News". *Digital Journalism*, p. 137–153.
25. Ellefsen, R. & Sandberg, S. (2022). "Everyday Prevention of Radicalization: The Impacts of Family, Peer, and Police Intervention." Studies in Conflict & Terrorism. https://www.doi.org/10.1080/1057610X.2022.2037185
26. Fabry, M., (2017). *Here's How the First Fact-Checkers Were Able to Do Their Jobs Before the Internet.* [Online] Available at: https://time.com/4858683/fact-checking-history/
27. Fact, F., (2020). *The challenges of online fact-checking*, London: Full Fact.
28. Festinger, Leon. (1962). A Theory of Cognitive Dissonance. Stanford, CA: Stanford University Press
29. Feldman L. (2017). *The Hostile Media Effect*. In: Kenski; Kate; Jamieson, Kathleen Hall. The Oxford Handbook of Political Communication. New York, Oxford University Press, 549–564. ISBN: 978-0-19-979347-1.
30. Fischer, S., (2020). *Fact-checking goes mainstream in Trump era.* [Online] Available at: https://www.axios.com/2020/10/13/fact-checking-trump-media
31. García-Marín, D. (2022). "Modelos algorítmicos y fact-checking automatizado. Revisión sistemática de la literatura." *Documentación de las Ciencias de la Información. Monográfico. Editorial Complutense*. ISSN-e: 1988-2890 https://dx.doi.org/10.5209/dcin.77472
32. García-Marín, D., (2020). Global infodemic: Information disorders, false narratives, and fact checking during the Covid-19 crisis. *Prisma Social.*
33. Graves, L. & Amazeen, M. A., (2019). Fact-Checking as Idea and Practice in Journalism. In: *Oxford research encyclopedia of communication.* Oxford: Oxford University Press.
34. Graves, L. & Cherubini, F., (2016). *The rise of fact-checking sites in Europe,* s.l.: The Reuters Institute for the Study of Journalism.
35. Graves, L. (2017). "Anatomy of a fact check: Objective practice and the contested epistemology of fact checking." *Communication, Culture & Critique*, 10:518–537. https://www.doi.org/10.1111/cccr.12163
36. Greitzer, F.L., Kuchar , O.A., and Huston , K. (2007) Cognitive science implications for enhancing training effectiveness in a serious gaming context. ACM J. Edu. Resources in Comput., Vol. 7, No. 3, Article 2 (August 2007), 10 pages. DOI=10.1145/1281320.1281322
37. Hameleers M, Powell TE, Van Der Meer TGLA, et al. (2020): "A Picture Paints a Thousand Lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media." *Political Communication*, 37(2): 281–301.

38. Herrero, E. & Herrera Damas, S., (2021). Spanish-Speaking Fact-Checkers around the world: profiles, similarities, and differences among fact checking professionals. *Revista de Comunicacion de la SEECI*, pp. 49-77.
39. Herrero, E. & Herrera Damas, S., (2021). Spanish-Speaking Fact-Checkers around the World: Profiles, Similarities, and Differences among Fact Checking Professionals. *Revista de Comunicación de la SEECI. 2021*, pp. 49-77.
40. Himma-Kadakas, M. & Ojamets, I., (2022). Debunking False Information: Investigating Jounalists' Fact-Checking Skills. *Digital Journalism*, 10(5), pp. 866-887.
41. Huynh, Viet-Phi & Papotti, Paolo (2019.) "A Benchmark for Fact Checking Algorithms Built on KnowledgeBases." CIKM '19, November 3–7, Beijing, China. https://doi.org/10.1145/3357384.3358036
42. Jarman, J. W. (2016) "Influence of political affiliation and criticism on the effectiveness of political fact-checking", *Communication Research Reports*, 33(1): 9-15. https://www.doi.org/10.1080/08824096.2015.1117436
43. Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Gonzalez, E. S. (2022). Understanding the adoption of Industry 4.0 technologies in improving environmental sustainability. Sustainable Operations and Computers.
44. Johnson, H.M., & Seifert, C.M. (1994). "Sources of the continued influence effect: When misinformation in memory affects later inferences." *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 20: 1420–1436.
45. Kan, I.P., Pizzonia, K.L., Drummey, A.B. et al.; Mikkelsen, E.J.E. (2021). "Exploring factors that mitigate the continued influence of misinformation." *Cogn. Research*. 6(76). https://doi.org/10.1186/s41235-021-00335-9
46. LaGarde, J. & Hudgins, D., (2018). *Fact vs. Fiction: Teaching Critical Thinking Skills in the Age of Fake News.* Portland, Oregon: International Society for Technology in Education.
47. Leiserowitz A.; Maibach E.; Rosenthal S.; Kotcher, J.; Carman, J.; Wang, X., Marlon, J.; Lacroix, K. & Goldberg, M. (2021). "Climate Change in the American Mind" in April 2021 Yale University and George Mason University; New Haven, CT: Yale Project on Climate Change Communication.
48. Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). "Misinformation and its correction: Continued influence and successful debiasing." *Psychological Science in the Public Interest*, 13(3): 106–131. https://doi.org/10.1177/1529100612451018
49. Lewandowsky, Stephan, and Sander Van Den Linden (2021). "Countering misinformation and fake news through inoculation and prebunking." *European Review of Social Psychology* 32 (2): 348-384.
50. Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E. Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., Vraga, E. K., Wood, T. J., Zaragoza, M. S. (2020). **The Debunking Handbook 2020.** Available at https://sks.to/db2020. DOI:10.17910/b7.1182.
51. Lukito J. (2020). "Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017." *Polit Commun*, 37(2): 238–255
52. Klein, H. A. (1974). Behavior modification as therapeutic paradox. American Journal of Orthopsychiatry, 44(3), 353–361. doi:10.1111/j.1939-0025.1974.tb00888.
53. MacDonald, Geoff, Nail Paul R, HAPER Jesse R (2010):" Do people use reverse psychology? An exploration of strategic self-anticonformity", in *Social influence. 6/2011, vol. 1, DOI: 10.1080/15534510.2010.517282*
54. Mahl, Daniela, Mike S. Schäfer, and Jing Zeng (2022):. "Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research." *New media & society* 14614448221075759. (online first)
55. Mantzarlis, A., (2015). *Introducing Poynter's International Fact Checking Network.* [Online]
56. Available at: https://www.poynter.org/news/introducing-poynters-international-fact-checking-network
57. Mantzarlis, A., (2018). Fact-checking 101. In: *Journalism, 'Fake News' & Disinformation. Handbook for Journalism Education and Training.* s.l.:UNESCO, pp. 81-95.
58. Marietta, M., Barker, D. C. & Bowser, T., (2015). Fact-Checking Polarized Politics: Does The Fact-Check Industry Provide Consistent Guidance on Disputed Realities?. *The Forum*, p. 577–596.

59. Margolin D.B.; Hannak A.; Weber I. (2018). “Political Fact-Checking on Twitter: When Do Corrections Have an Effect?.” *Political Communication.* 35(2): 196–219
60. Marwick, E. W. (2018). “Why do people share fake news? A sociotechnical model of media effects.” *Georgetown Law Technologic Review*, 474.
61. Miranda, S., Nogueira, D. & Mendes, A. (2019). “Automated Fact Checking in the News Room,” WWW ’19, May 13–17-, 2019, San Francisco, CA, USA, 2019 IW3C2 (International World Wide Web Conference Committee), published ACM ISBN 978-1-4503-6674-8/19/05.
62. Nail, PR and Van Leeuwen, MD. (1993). An analysis and restructuring of the diamond model of social response. *Personality and Social Psychology Bulletin*, 19: 106–116.
63. Nieminen, S. & Rapeli, L., (2018). Fighting Misperceptions and Doubting Journalists'Objectivity: A Review of Fact-checking Literature. *Political Studies Review*, pp. 1-14.
64. Nyhan, B., Porter, E., Reifler, J. & Wood, T. J., (2020). Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability. *Political Behavior*, p. 939–960.
65. Nyhan, B., & Reifler, J. (2010). “When corrections fail: The persistence of political misperceptions.” *Political Behavior*, 32(2):303–330. https:// doi.org/10.1007/s11109-010-9112-2
66. Nyhan B., Reifler, J., Ubel P. (2013). “The Hazards of Correcting Myths about Health Care Reform.” Medical care, 51(2): 127-132.https://www.doi.org/ 10.1097/MLR.0b013e318279486b
67. Nyhan, B. (2021). “Why the backfire effect does not explain the durability of political misperceptions.” *Proceedings of the National Academy of Sciences*, 18(15), https://doi.org/10.1073/pnas.1912440117
68. Ornebring, H. & Mellado, C., (2016). Valued Skills among Journalists: An exploratory comparison of six European nations. *Journalism*, 19(4), pp. 1-19.
69. Pal, A. & Banerjee, S., (2019). Understanding online falsehood from the perspective of social problem. In: *Handbook of Research on Deception, Fake News, and Misinformation Online.* Hershey, PA: IGI Global, pp. 1-17.
70. Pemmet, James & Anneli Kimber Lindwall (2021) *Fact-Checking and Debunking. A Best Practice Guide to Dealing with Disinformation*, published by NATO Strategic Communications Center of Excellence.
71. Pennycook, Gordon, and David G. Rand (2021):. "The psychology of fake news." *Trends in cognitive sciences* 25 (5): 388-402.
72. Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4944-4957.
73. Perez-Escolarr, M., Ordonez-Olmedo, E. & Alaide-Pulido, P., (2021). Fact-Checking Skills And Project-Based Learning About Infodemic And Disinformation. *Thinking Skills and Creativity*, pp. 1-11.
74. Persily, N., (2017). The 2016 U.S. Election: Can Democracy Survive the Internet?. *Journal of Democracy, Volume 28, Number 2,* pp. 63-76.
75. Poynter, n.d. *International Fact-Checking Network. Empowering fact-checkers worldwide.* [Online] Available at: https://www.poynter.org/ifcn/about-ifcn/
76. Roozenbeek, J., & van der Linden, S. (2019a). "The fake news game: Actively inoculating against the risk of misinformation". *Journal of risk research*, 22(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491
77. Roozenbeek, J., & van der Linden, S. (2019b). "Fake news game confers psychological resistance against online misinformation". *Nature Humanities and Social Sciences Communications*, 5(65). https://doi.org/10.1057/s41599-019-0279-9
78. Roozenbeek, J., & van der Linden, S. (2020). "Breaking Harmony Square: A game that “inoculates” against political misinformation". *The Harvard Kennedy School Misinformation Review*, 1(8). https://doi.org/10.37016/mr-2020-47
79. Roozenbeek, J., & Van Der Linden, S. (2022). How to combat health misinformation: A psychological approach. *American journal of health promotion*, *36*(3), 569-575.
80. Shifman, L., (2013). Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer Mediated Behavior*, p. 362–377.

81. Stencel, M., Ryan, E. & Luther, J., (2022). *Fact-checkers extend their global reach with 391 outlets, but growth has slowed.* [Online] Available at: https://reporterslab.org/tag/fact-checking-census/
82. Stenger, M., (2016). *8 Ways to Hone Your Fact-Checking Skills.* [Online] Available at: https://www.opencolleges.edu.au/informed/features/8-ways-to-hone-your-fact-checking-skills/
83. Susmann M.K. & Wegener, D.T. (2022). "The role of discomfort in the continued influence effect of misinformation." *Memory & Cognition,* 50: 435–448, https://doi.org/10.3758/s13421-021-01232-8
84. Swire, Briony, et al (2017):. "Processing political misinformation: Comprehending the Trump phenomenon." *Royal Society open science* 4.3 160802.
85. Traberg CS, Roozenbeek J, van der Linden S (2022) "Psychological Inoculation against Misinformation: Current Evidence and Future Directions" *The ANNALS of the American Academy of Political and Social Science*. 700(1):136-151. doi:10.1177/00027162221087936
86. Trevors, G.J. (2022). "The Roles of Identity Conflict, Emotion, and Threat in Learning from Refutation Texts on Vaccination and Immigration." *Discourse Processes,* 59(1-2): 36-51. https://doi.org/ 10.1080/0163853X.2021.1917950.
87. Valkenburg P.M., Oliver M.B. (2020). *Media Effects Theories: An Overview*. In: Oliver, Mary Beth; Raney, Arthur A.; Bryant, Jennings. Media Effects: Advances in Theory and Research. Fourth Edition. Routledge Communication Series, Kindle Edition.
88. Van der Bles AM, Van der Linden S, Freeman ALJ, et al.(2020). "The effects of communicating uncertainty on public trust in facts and numbers." *Proc Natl Acad Sci U S A*, 117(14): 7672–7683
89. van der Linden S, Roozenbeek J and Compton J (2020) "Inoculating Against Fake News About COVID-19". *Front. Psychol.* 11:566790. doi: 10.3389/fpsyg.2020.566790
90. van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). "Inoculating the public against misinformation about climate change". *Global Challenges*, 1(2): 1-17 10.1002/gch2.201600008
91. Vosoughi, S., Roy, D., & Aral, S. (2018). "The spread of true and false news online". *Science*, 359(6380): 1146–1151.
92. Walter, N.; Cohen,J.; Holbert R.L & Morag, Y. (2019). "Fact-Checking: A Meta-Analysis of What Works and for Whom." *Political Communication,* https://www.doi.org/10.1080/10584609.2019.1668894
93. Walter, N. & Murphy, S. T. (2018). "How to unring the bell: A meta-analytic approach to correction of misinformation." *Commun*. 85: 423–441
94. Willis, RH. 1965. Conformity, independence, and anticonformity. *Human Relations*, 18: 373–388.
95. Wintersieck, A. L. (2017). "Debating the Truth: The Impact of Fact-Checking During Electoral Debates." *American Politics Research*, 45(2): 304–331. https://doi.org/10.1177/1532673X16686555
96. Yaqub, W. et al., (2020). Effects of Credibility Indicators on Social Media News Sharing Intent. s.l., Publication History, pp. 1-14.
97. *RESIST 2. Counter disinformation toolkit*, UK Government Communication Service, 2021.
98. *Journalism, 'Fake News' & Disinformation Handbook for Journalism Education and Training*, 2018 (available at Journalism, fake news & disinformation: handbook for journalism education and training - UNESCO Digital Library);
99. *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression;* Broadband Commission research report on 'Freedom of Expression and Addressing Disinformation on the Internet, 2020, available at Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression (broadbandcommission.org)
100. *MSB Countering information influence activities A handbook for communicators*. Swedish Civil Contingencies Agency (MSB), 2019 (available at Countering information influence activities : A handbook for communicators (msb.se))
101. World Health Organization, 2022. Infodemic. [Online] Available at: https://www.who.int/health-topics/infodemic#tab=tab_1

4. LEGAL FRAMEWORK FOR COUNTERING DISINFORMATION AND PROPAGANDA - THE WAY FORWARD

Introduction

Chapter 4 looks at the legal framework for countering disinformation and propaganda. The first section analyses the legal frameworks of Malta, Spain and Romania and examines how freedom of expression has been affected by disinformation and how the aforementioned states address disinformation through their legal frameworks. After a general overview of the aforementioned legal frameworks, the second section examines the role of data protection in combating disinformation. It identifies different dimensions of data protection that may be relevant for preventing disinformation, such as profiling, automated decision-making, the principles of data protection by design/data protection by default and sensitive personal data. The third section looks at the relevant national and European case law, where available, to see how the courts interpret or in some cases fill in the gaps in the area of fake news and disinformation. Finally, the last section of this chapter examines some of the most commonly used technological tools to detect and prevent disinformation and analyses their potential and limitations.

Digital competences addressed

2.3 Engaging citizenship through digital technologies

4.1 Existing Legislation and Intersection with Media Freedom

Ana Ćuća, Aitana Radu

Abstract

The first section of the section is dedicated to a short introduction into the right to freedom of expression, and how this right is protected in national and international legislation. The section then continues with an analysis of how freedom of expression is connected to the spread of disinformation, which is focused on three main case studies: Malta, Spain and Romania. The objective is to examine recent legislative changes aimed at addressing disinformation and whether these changes are in compliance with human rights standards. To illustrate how these legislative changes work in practice, the deliverable focuses on the case of the COVID-19 pandemic and the disinformation surrounding the measures aimed to control the pandemic, which was a common challenge for all three countries examined. Lastly, this section will look into new legal developments, related to the role of intermediaries in spreading disinformation. The objective of this section is to provide an overview of relevant legislation in relation to disinformation. To do so the section focuses on three different legal systems: Malta, Spain and Romania to showcase both common approaches but also national differences in tackling disinformation. Moreover, the section employs the example of disinformation related to the COVID-19 pandemic measures to illustrate how legislative provisions have been applied in practice.

Main research questions addressed

- What is the connection between freedom of expression and the fight against disinformation?
- How do the legal frameworks of Malta, Spain and Romania protect freedom of expression and address disinformation?
- Do countries recognize disinformation as a hybrid threat and have they addressed the COVID-19 Infodemic?
- What is the purpose of the Digital Services Act?

Constitutional overview

Each state must safeguard freedom of expression while at the same time ensuring that the dissemination of fake news is accurately tackled. Article 11 of the Charter of Fundamental Rights of the European Union (2000/C 364/01) defines freedom of expression and information as: “[the] right [that] includes freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers” (Charter of Fundamental Rights of the European Union, Article 11). Similarly, Article 10 of the European Convention on Human Rights states:

1. Everyone has the right to freedom of expression. This right shall include the freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.
2. The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary (Charter of Fundamental Rights of the European Union, Article 10).

The right to freedom of expression is universal. Therefore, states must protect and ensure that their legal systems provide adequate and effective safeguards for protecting freedom of expression. These safeguards are usually defined in the constitution, given that the protection of freedom of expression is essential for the democratic political process and development. Freedom of expression and information is addressed in constitutions, either in combination with other fundamental rights and freedoms or as a separate fundamental human right. The following paragraphs will show how the constitutions of Malta, Spain and Romania, the countries analysed in this chapter, follow the standards set by the Charter of Fundamental Rights of the European Union, and the European Convention on Human Rights.

a. Freedom of expression as one of the fundamental rights and freedoms of the individual

The Maltese Constitution does not have a separate freedom of expression clause. Freedom of expression is protected under a clause referring to the fundamental rights and freedoms of the individual. Article 32 of the Constitution of Malta reads as follows:

> Whereas every person in Malta is entitled to the fundamental rights and freedoms of the individual, that is to say, the right, whatever his race, place of origin, political opinions, colour, creed, sex, sexual orientation or gender identity, but subject to respect for the rights and freedoms of others and for the public interest, to each and all of the following, namely
> (a) life, liberty, security of the person, the enjoyment of property and the protection of the law;
> (b) freedom of conscience, of expression and of peaceful assembly and association; and
> (c) respect for his private and family life,
> the subsequent provisions of this Chapter shall have effect for the purpose of affording protection to the aforesaid rights and freedoms, subject to such limitations of that protection as are contained in those provisions being limitations designed to ensure that the enjoyment of the said rights and freedoms by any individual does not prejudice the rights and freedoms of others or the public interest (Constitution of Malta, Article 32).

b. Freedom of expression linked with production, academic freedom, right to receive truthful information

Spain sets protection safeguards for the freedom of expression in Article 20 of its Constitution. It links the right to freely express and share thoughts with the right to literary, artistic, scientific and technical production; the right to academic freedom and; the right to freely communicate and receive truthful information. The following safeguards apply to all rights mentioned:

> 2. The exercise of these rights may not be restricted by any form of prior censorship.
>
> 3. The law shall regulate the organization and parliamentary control of the mass communication means under the control of the State or any public agency and shall guarantee access to such means by significant social and political groups, respecting the pluralism of society and of the various languages of Spain.
>
> 4. These freedoms are limited by respect for the rights recognized in this Part, by the legal provisions implementing it, and especially by the right to honour, to privacy, to the own image and to the protection of youth and childhood.
>
> 5. The seizure of publications, recordings and other means of information may only be carried out by means of a court order (Spanish Constitution, Article 20).

c. Freedom of expression as a separate fundamental right

In contrast to the Maltese and Spanish examples, Article 30 of the Romanian Constitution separately addresses freedom of expression, defining safeguards and limitations to this right:

> (1) Freedom of expression of thoughts, opinions or beliefs and freedom of creations of any kind, through live speech, writing, images, sounds or other means of public communication, is inviolable.
>
> (2) Censorship of any kind is prohibited.
>
> (3) Freedom of the press also implies the freedom to establish publications.
>
> (4) No publication may be suppressed
>
> (5) The law may impose on mass media the obligation to make public the source of funding.
>
> (6) Freedom of expression cannot prejudice the dignity, honour, private life of the person, nor the right to one's own image.
>
> (7) Defamation of the country and the nation, incitement to war of aggression, national, racial, class or religious hatred, incitement to discrimination, territorial separatism or public violence, as well as obscene manifestations contrary to good morals, are prohibited by law.
>
> (8) The civil liability for the information or for the creation brought to public knowledge rests with the editor or producer, the author, the organizer of the artistic manifestation, the owner of the means of multiplication, of the radio or television station, in accordance with the law. Crimes that can be conducted through the press will be established by law (Constitution of Romania, Article 30).

Freedom of expression underpins other human rights; thus, it does not come as a surprise that some states in their constitutions connect it with other fundamental rights. As freedom of expression is a component of other fundamental rights, clear safeguards must be set, as well as restrictions or penalties. According to the standards set by the European Convention on Human Rights, any restrictions to freedom of expression must be prescribed by law and must be necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary (Bychawska-Siniarska, 2017). Just as there are differences in introducing the freedom of expression in their constitutions, states can also limit freedom of expression differently:

a. Limiting freedom of expression to protect human dignity, privacy or personal image

An example of this approach is the Romanian Constitution which clearly states that “freedom of expression cannot prejudice the dignity, honour, private life of the person, nor the right to one's own image” (Constitution of Romania, Article 30).

b. Limiting freedom of expression to protect youth and children

Article 20 of the Spanish Constitution sets restrictions to the freedom of expression, to protect “the right to honour, to privacy, to the own image and the protection of youth and childhood (Spanish Constitution, Article 20).”

c. Limiting freedom of expression by consent or as a parental measure

Article 41 of the Maltese Constitution introduces the possibility of limiting freedom of expression by consent or as part of the so-called parental discipline:

> Except with his own consent or by way of parental discipline, no person shall be hindered in the enjoyment of his freedom of expression, including freedom to hold opinions without interference, freedom to receive ideas and information without interference, freedom to communicate ideas and information without interference (whether the communication be to the public generally or to any person or class of persons) and freedom from interference with his correspondence (Constitution of Malta, Article 41).

States have the responsibility to enact laws and regulations that need to create an “enabling” environment for individuals to exercise their right to freedom of expression. The legislative framework needs to be carefully assessed, so as not to hinder the exercise of this right. Special focus must be put on laws that could be deterring, such as those that define legal liability. Whereas these can help in tackling the issue of disinformation, they can foster censorship, self-censorship, and criminal, financial and administrative sanctions. In order to protect freedom of expression, some countries, like Spain, opted for full prohibition of censorship. Article 20 of the Spanish Constitution stipulates that the freedom of expression may not be restricted “by any form of censorship” (Spanish Constitution, Article 20). Although full prohibition of censorship might seem as a way for the state to guarantee no limitations will be set to individual freedom of expression or to suppress the opposing views, such an approach may have its shortcomings. By avoiding censorship, countries with the same approach fail to address the spread of illegal content which can increase the need for more pro-active actions (European Commission, 2017). Incitement to terrorism, xenophobic and racist speech that publicly incite hatred and violence, as well as child sexual abuse materials should be adequately flagged and removed. In comparison with Spain, Romania prohibits censorship with several exceptions. While article 30 of the Romanian Constitution states that "any censorship shall be prohibited" paragraph 7 of the same article stipulates (Constitution of Romania, Article 30):

> Any defamation of the country and the nation, any instigation to a war of aggression, to national, racial, class or religious hatred, any incitement to discrimination, territorial separatism, or public violence, as well as any obscene conduct contrary to morality shall be prohibited by law (Constitution of Romania, Article 30).

Article 30 of the Romanian Constitution serves as an example of a legal clause that successfully integrated both the need to protect freedom of expression by limiting censorship, while at the same time acknowledging the importance of addressing illegal content. In comparison with two previous examples, the Maltese Constitution does not have any references to censorship. Whatever approach states opt to take on when it comes to limiting censorship, legal clauses referring to censorship must be written in a way that does not restrict freedom of expression. If there are some exceptions, these exceptions must be precisely defined, so there is control over the scope of restrictions exercised by public authorities. Among the different forms of interference, censorship before publishing can be the most dangerous, as it stops the transmission of information and ideas to those who want to receive them.

Criminal convictions and sentences are one of the most dangerous post-expression interferences with the freedom of expression. Similarly, in the case of censorship, legal liability clauses need to be written in a way where the scope of the restriction is clear, minimising legal uncertainty. An example of the legal liability clause is Article 30, paragraph 8 of the Romanian Constitution which reiterates:

> Civil liability for any information or creation made public falls upon the publisher or producer, the author, the producer of the artistic performance, the owner of the copying facilities, radio or television station, under the terms laid down by law. Indictable offences of the press shall be established by law (Constitution of Romania, Article 30).

States need to safeguard freedom of expression while, at the same time, ensuring that the dissemination of fake news is accurately tackled. To deter people/entities from sharing disinformation, some states have introduced legal provisions that ensure the right of an individual to receive truthful information. These provisions, apart from ensuring individuals' rights, also prescribe responsibilities to media who produce and share disinformation. Article 31, of the Romanian Constitution, serves as an example of such a legal provision:

> (1) The right of the person to have access to any information of public interest cannot be restricted.
>
> (2) In their area of responsibility, public authorities are obliged to ensure the correct information of citizens on public affairs and issues of personal interest.
>
> (3) The right to information cannot go against measures to protect young people or national security.
>
> (4) The mass media, public and private, are mandated to ensure correct information of public opinion.

(5) Public radio and television services are autonomous. They must guarantee the right of important social and political groups to express their views. The organization of these services and the parliamentary control over their activity will be regulated by organic law (Constitution of Romania, Article 31).

Lack of precision in provisions which assign responsibility for spreading disinformation raises a question – what are the entities included under the "public and private media"? Who is to be held accountable? These questions are important in the context of both pre and post-expression interference. As a valuable starting point, both in the context of legal liability and, more so, of protection standards, the Council of the European Union, advises that "efforts to protect journalists should not be limited to those formally recognised as such, but should also cover support staff and others, such as "citizen journalists", bloggers, social media activists and human rights defenders, who use new media to reach a mass audience" (Council of the European Union, 2014). The recommendations made by the Council of the European Union are indicative of the changing times in terms of media protection and the acknowledgement that different and, perhaps, stronger safeguards need to be put into place at national and European levels. In an increasingly volatile and polarised world, often combined with a tendency of both democracies and authoritarian regimes to resort to coercion, it is important to protect journalists, "citizen journalists", human rights defenders. Although the Guidelines issued by the Council of the European Union are not legally binding, by following them, the EU Member States are showing political commitment to protect and advance the work of journalists, human rights defenders and citizens, allowing them to peacefully stand up against any unfairness.

The national response to disinformation

The European Commission has emphasized that the "primary obligation of state actors in relation to freedom of expression and media freedom is to refrain from interference and censorship and to ensure a favourable environment for inclusive and pluralistic debate" (European Commission, 2018). However, with the increasing use of disinformation to undermine democracies, the European Commission has changed its approach, advocating for national authorities to implement a variety of measures that would adequately respond to the new challenges posed by disinformation, while, at the same time, protecting freedom of expression. The Cambridge Analytica scandal showed how disinformation can be used to target voters with individually-tailored content, which is adjusted in real-time to reflect the debate that develops around critical electoral issues. Disinformation was not only used to undermine elements of good democracy, but it also destroyed trust in mainstream media, allowing alternative news to flourish.

As a first step to addressing the issue of disinformation, the European Commission had to decide what falls under this term. The European Commission defined 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 includes threats to democratic processes as well as to public goods such as Union citizens' health, environment or security" (European Commission, 2018b).

Although an EU-wide unified response to the issue of disinformation was expected, EU Member States opted for an individual approach, given that disinformation is harmful, but still not illegal per se. EU legislation currently differentiates illegal content from harmful content. Incitement to terrorism, xenophobic and racist speech that publicly incites hatred and violence, as well as child sexual abuse materials are illegal in the EU (European Commission, 2017). Harmful content refers to information that does not strictly fall under legal prohibitions but that might nevertheless have harmful effects, inter alia, disinformation. Whereas with illegal content the EU law is clear, in the case of harmful but legal content, the situation becomes more complicated as it consists of information that may be inadequate, but whose legality varies significantly across Member States.

Consequently, some Member States decided to address harmful content, by introducing in their Criminal Code legal provisions that define the act of producing and sharing disinformation and prescribe punitive measures. Malta serves as an example of the EU Member State, which has legal provisions that address the issue of producing and sharing disinformation. Article 82 of Malta's Criminal Code stipulates the following:

> Whosoever shall maliciously spread false news which is likely to alarm public opinion or disturb public good order or the public peace or to create a commotion among the public or among certain classes of the public, shall, on conviction, be liable to imprisonment for a term from one to three months:
>
> Provided that if any disturbance ensues in consequence of the offence, or if the offence has contributed to the occurrence of any disturbance, the offender shall be liable to imprisonment for a term of not less than one month but not exceeding six months and to a fine(multa) not exceeding one thousand euro (€1,000) or both such fine and imprisonment (Malta's Criminal Code, Article 82).

Article 82 of Malta's Criminal Code shows significant similarities with the European Commission's definition of disinformation. It highlights false information that is shared with malicious intentions, intending to deceive and harm the public. However, in comparison with the Commission's definition, no financial gain is needed for the malicious news to be labelled as false or disinformation.

Alongside the example of Malta, there are other Member States who have also introduced into their Criminal Codes provisions which address the intentional spread of disinformation and these are France, Croatia, Greece, Slovakia, Czech Republic, and Cyprus (Fathaigh, R., et al., 2021). In addition to those, others have decided to take a harder stance in addressing the issue of disinformation. For example, Romania followed the same approach as the before-mentioned countries, with one difference – it interlinked the spread of disinformation with the threats to national security. Precisely, Article 404, of the Romanian Criminal Code condemns the knowing spread of false information if it threatens national security, and establishes a sentence of between 1 and 5 years (Romanian Penal Code, Article 404). National security is defined by the Law on National Security (51/1991) as:

> … a state of social, economic and political legality, equilibrium and stability that is necessary to the existence and development of the Romanian national state - a sovereign,

unitary, independent and indivisible state, to the maintenance of legal order as well as of the climate for the unhampered exercise of the fundamental rights, freedoms and duties of the citizens, in accordance with the democratic principles and rules provided by the Constitution (Romanian Law on National Security, Article 1).

Since the current definition of national security does not precisely define which acts might pose a threat, there is a question of where the authorities draw the line and how they determine whether the spreading of (specific) false information is a threat to national security. The lack of a clear distinction between what falls under the scope of this legal provision raises legal uncertainty. Enjoyment of the right to freedom of expression may be limited, however, such limitations must be precisely defined, otherwise, they constitute arbitrary and discriminatory influence.

Similarly to the Romanian example, Spain also approached disinformation as a possible threat to national security. In 2019, Spain introduced disinformation in their *National Cybersecurity Strategy*, arguing that “malicious use of personal data and disinformation campaigns have high potential to destabilise society” (Spanish Government, 2019). As a result of recognising disinformation as a cyber threat, the CCN-Cert, an organisation of Spanish intelligence has become involved in the fight against disinformation, which is understood as part of cyber defence in the broadest sense. The CCN-Cert operates a national security centre whose aim is to achieve safer and more reliable cyberspace, preserving classified and sensitive information. According to the Regulation and Law on the Public Sector Legal Regime Currently, CCN-Cert is responsible for the management of cyber-incidents affecting any public body or company. Currently, Spanish Criminal Code doesn’t have any punitive measures prescribed for spreading disinformation (Centro Criptologico Nacional, 2019). However, just a year later after integrating disinformation into the National Cybersecurity Strategy and nominating CCN-Cert to monitor the issues of disinformation, Spain introduced their *Procedure of Action against Disinformation* that was approved by the Department of Homeland Security (DSN). The Procedure of Action against Disinformation contains four levels of action:

Level I. 1. Monitorization and surveillance: detection, early warning, notification, and analysis; 2. Participation in the European Union´s Rapid Alert System (RAS) and activation of protocols; 3. Research the possible origin, the purpose and tracking of its activity; 4. Deciding if the event is elevated to a higher body or if it is finished.

Level II. 1. Call, tracking and evaluation of the alert by the Permanent Commission against disinformation; 2. Analysis of the situation and support for the definition of proposals for action; 3. Activating, where appropriate, a Coordination Cell against Disinformation activated ad hoc by the Director of the Department of Homeland Security; 4. Decision on its elevation or the carrying out of a public communication campaign led by the State Secretariat for Communication depending on the nature of the disinformation campaign.

Level III. 1. Information at the political-strategic level by the Secretary of State for Communication; 2. Monitoring and evaluation of the alert by the Situation Committee or Public Communication agreed according to guidelines of the Situation Committee.

Level IV. 1. Coordination of the response at the political level by the National Security Council in case of public attribution of a disinformation campaign to a third State (Spanish Government,2020).

The relationship between disinformation and human rights is double-edged. Disinformation infringes a range of core rights. These include the freedom of thought; the right to privacy; the right to participation. Disinformation also weakens democracies, it has the potential to interfere with elections, and can feed digital violence. However, counter-disinformation initiatives also carry risks for human rights and democracy. In some states measures against disinformation have constricted human rights. That is why it is important to find appropriate ways for legislative and executive bodies to regulate the spread of disinformation, while at the same time being attentive to ways these may impact human rights.

Disinformation – Use Cases

In recent years, the transmission of disinformation has increased dramatically across the world. Recently disinformation has mostly been mentioned in the context of the exposure of Russian interference in the American elections and the war in Ukraine, both cases illustrating how concerted disinformation campaigns can foster democratic regression and promote authoritarian regimes. The COVID-19 pandemic has also led to a spike in disinformation, through the online dissemination of pseudoscience, and conspiracy theories, thereby, causing distrust in public institutions and risking people’s lives (e.g. through their refusal to get vaccinated and/or follow other health measures). The following subchapters will provide examples of how the three countries under analysis, namely Malta, Romania and Spain have responded to the newest challenges linked to disinformation.

a. Disinformation as a Hybrid Threat

In light of Russian interference in the American elections, and the war in Ukraine, the European Commission issued a *Joint Framework on countering hybrid threats*, where it reiterated its position that “while definitions of hybrid threats vary and need to remain flexible to respond to their evolving nature, the concept aims to capture the mixture of coercive and subversive activity, conventional and unconventional methods… Massive disinformation campaigns, using social media to control the political narrative or to radicalise, recruit and direct proxy actors can be vehicles for hybrid threats” (European Commission, 2016).

Keeping in mind the central role of disinformation in the destabilization of democratic countries, some EU member states have decided to integrate hybrid threats in their defence strategies, while others decided to follow the Commission’s Framework, while gradually integrating the Commission’s recommendations in their policy framework.

Romania is an example of a country that took a more proactive role in detecting and protecting from hybrid threats, which include disinformation. The *Romanian Strategy for National Defence*, adopted in 2020 connects hybrid threats and disinformation, as one of the manifestations of these threats. According to the defence strategy, hybrid threats are a multi-faceted form of attack, of which disinformation is one, aiming to weaken the population's belief in the measures taken by the state. These are addressed in the section on threats to and vulnerabilities of the Romanian state:

> 126. Hostile actions aimed to influence the public, change perceptions and influence the behaviour of civil society, constitute a constant threat to social security, and can potentially increase given the diversification of means of communication in the online environment.
>
> 158. The persistence of some legislative gaps in the field of national security or in terms of countering informational aggressions, respectively at the level of regulating the tools necessary to prevent and counter-propaganda with destabilizing purposes, including in the event of hybrid campaigns (Romanian Presidential Administration, 2020).

Spain's Strategy for National Defence mentions the existence of hybrid threats, but the emphasis is on the use of hybrid strategies to respond to ever-evolving threats, some of which are the information shared to destabilize the society:

The use of hybrid strategies combining conventional and asymmetric procedures leads to a framework of intense confrontation in cyberspace and the information environment. The use of force goes hand in hand with psychological campaigns designed to discredit our actions and spread confusion in public opinion. In the cyberspace and information fields, it is usual for some adversaries to hide their actions and apply their strategies in a grey zone, located below what has been identified as our response threshold.

Malta recognizes the role of disinformation in the context of hybrid threats but does not refer to it in its National Defense Strategy. Moreover, in the case of Malta, the connection between disinformation and hybrid threats was emphasized mainly in the context of the COVID-19 pandemic. This being said, the Maltese Ministry for Foreign and European Affairs has acknowledged and endorsed the threat posed by disinformation and hybrid threats and has been participating in coordinated and comprehensive cross-administrative discussions on hybrid threats to ensure that EU Member States benefit from cooperation within the Union as much as possible and to improve their capacity to combat hybrid threats and disinformation.

b. Disinformation and the Infodemic

The COVID-19 crisis showed the extent of the threat disinformation poses to our society. In the words of the European Commission "the infodemic - the rapid spread of false, inaccurate or misleading information about the pandemic – has posed substantial risks to personal health, public health systems, effective crisis management, the economy and social cohesion" (European Commission, 2021, 1). In light of the new circumstances, the Commission published its *Guidance on Strengthening the Code of Practice on Disinformation* (European Commission, 2021). As per Commission's view, relevant stakeholders (social media and other digital platforms) must "step up their measures to address gaps and shortcomings in the Code and create a more transparent, safe and trustworthy online environment" (European Commission, 2021, 2) (see also Chapter 6 for more information on this document).

While some countries decided to rely on informative campaigns and publicly assign the responsibility to individuals to be mindful when sharing content related to COVID-19, others resorted to more punitive measures.

Malta, for example, relied on a more holistic approach when addressing the infodemic. In a press release issued by the Government, the responsibility for stopping disinformation was allocated to individuals:

> Undoubtedly, the spread of disinformation around COVID-19 has led to potential harmful consequences that could contribute to potential destabilisation, if not cripple, whole societies. Understanding threats, developing responses and sharing insights, are now key to finding comprehensive approaches to tackle hybrid threats and disinformation.
>
> False information, mistrust, and panic will undoubtedly keep increasing, unless proper counter measures are enacted. The responsibility is on us citizens, to counter the further proliferation of disinformation, by simply thinking before clicking and reflecting on possible consequences before sharing information on today's social media platforms (Faruggia, 2020).

Spain took a similar approach to the one employed by Malta. The Spanish National Police issued a guide to prevent citizens from being manipulated by disinformation, by compiling main narratives that are built on disinformation and fake news (Torres et al., 2021).

By comparison to Malta and Spain, Romania took a harder and more systematic stand when tackling the infodemic. Romania declared a state of emergency on the 16th of March 2020, which was in force until the 15th of May of the same year. Between May 2020 and March 2022, Romania maintained a ”state of alert”, which included less severe restrictions. The legislation that expressly allowed authorities to forbid content online and to eliminate websites, namely Article 54 of the Decree 195, 16 of March 2020 was in force during the two months of the state of emergency and was included in the decree that introduced the state of emergency. The article states that:

> (1) Public institutions and authorities, as well as private operators contribute to the public information campaign regarding the measures adopted and the activities carried out at the national level.
>
> (2) In the event of the propagation of false information in the mass media and in the online environment regarding the evolution of COVID-19 and the protection and prevention measures, institutions and public authorities take the necessary measures to inform the population correctly and objectively in this context.
>
> (3) Hosting service providers and content providers are obliged to, on the reasoned decision of the National Authority for Administration and Regulation in Communications, to immediately interrupt, after notifying the users, the transmission through an electronic communications network or the storage of the content, by eliminating it at the source, if the respective content promotes false news regarding the evolution of COVID-19 and to protection and prevention measures.
>
> (4) In the situation where the elimination at the level of the source of the content provided for in par. (3) is not feasible, the providers of electronic communications networks intended for the public are mandated, upon the reasoned decision of the National Authority for Administration and Regulation in Communications, to immediately block access to said content and to inform the users.

> (5) Upon the reasoned decision of the National Authority for Administration and Regulation in Communications, providers of electronic communications networks intended for the public have the obligation to immediately block the access of users in Romania to content that promotes fake news regarding the evolution of COVID-19 and to protection and prevention measures and it is transmitted in an electronic communications network by the persons from para. (3) which is not under the jurisdiction of national law (Romanian Presidential Administration, 2020).

Based on the decree, the National Authority for Administration and Regulation in Communications was able to close down 15 websites accused of spreading COVID 19-related disinformation, through individual decisions. The decisions lapsed after the lifting of the state of emergency in May 2020 (ANCOM, 2020).

Digital Services Act – a step towards regulating disinformation in digital space

Digital technologies, business models and services have changed at an unprecedented pace. Having experienced the impact of digital technologies used as an intermediary for sharing disinformation to destabilize democracies and political systems, EU Member States are increasingly introducing, or are considering introducing, national laws that strive to regulate digital space. Russian intervention in the American elections, the infodemic and the current war in Ukraine, once again showed why it is important to build a common legal framework that would adequately and timely respond to new challenges.

With its new *Digital Services Act* (2022), the European Union took an important step to ensure a safer online environment. The DSA protects the digital space against the spread of illegal content, disinformation and ensures the protection of users' fundamental rights. The DSA defines clear responsibilities and accountability for providers of intermediary services, such as social media, online marketplaces, very large online platforms (VLOPs) and very large online search engines (VLOSEs). The rules are designed asymmetrically, which means that larger intermediary services with significant societal impact (VLOPs and VLOSEs) are subject to stricter rules. According to the Act:

> When recipients of the service are presented with advertisements based on targeting techniques optimised to match their interests and potentially appeal to their vulnerabilities, this can have particularly serious negative effects. In certain cases, manipulative techniques can negatively impact entire groups and amplify societal harms, for example by contributing to disinformation campaigns or by discriminating against certain groups. Online platforms are particularly sensitive environments for such practices and they present a higher societal risk (Digital Services Act, paragraph 69).

Under the DSA, platforms will not only have to be more transparent but will also be held accountable for their role in disseminating illegal and harmful content. Specifically:

- Wide-ranging transparency obligations regarding the steps taken by platforms to combat illegal and fake information.
- Measures to counter the sale of illegal goods/services.
- Transparency obligations concerning online advertisements.

- Updated liability regime for online intermediaries
- A new crisis response mechanism (in the context of the ongoing situation in Ukraine and the potential impact of the manipulation of online information) which will facilitate the analysis of the impact of the activities of VLOPs/VLOSEs on the crisis and rapidly decide on proportionate and effective measures to implement to protect fundamental rights.
- Platforms accessible to minors must implement special protection measures to ensure their online safety and will be prohibited from using targeted advertising based on the use of minors' personal data.
- Restrictions on targeted advertising based on profiling using special categories of personal data such as sexual orientation or religious beliefs and the use of "dark patterns" on the interface of online platforms.

VLOPs/VLOSEs must also:

- Analyse the systemic risks their platforms create and implement effective content moderation mechanisms to address them.
- Provide transparency on the key parameters of decision-making algorithms used to provide content and offer users a system for recommending content which is not based on profiling.
- There are heavy fines for non-compliance of up to 6% of annual global income/turnover (Trynor, 2022).

DSA takes an expansive view of digital regulatory policy by proposing to introduce legally binding tools, especially with regard to the accountability and transparency of digital platforms. These measures seek to enhance the EU's democratic resilience and regulatory toolbox. The DSA rules entered into force on the 16th of November. If successfully implemented across the EU, they will present a ground-breaking horizontal regulation that will mark the beginning of a new relationship between online platforms, users and regulators in the European Union and beyond.

A fully comprehensive approach must be taken to counter disinformation effectively. A series of instruments must be developed both internationally (EU-wide) and nationally. While developing external actions against disinformation stronger human rights and democracy considerations must be made. Response to disinformation is needed on several levels:

Laws and Regulations

- All restrictions to freedom of expression must be clear and must respect the principle of legal certainty.
- Legislation restricting the right to freedom of expression must be applied by the body which is independent of political or private influence.
- Remedy measures against the abusive application of legislation which limits freedom of expression must be available.
- States must take care to ensure that anti-terrorism laws, treason laws or similar provisions relating to national security must be applied in a manner that agrees with their obligations under international human rights law.

Media

- Journalists, including freelance journalists, media actors and individuals must be committed to producing quality journalism, and have access to life-long training opportunities to update their skills and knowledge, specifically concerning their duties and responsibilities in the digital environment.
- Media must develop effective self-regulatory mechanisms to deal with disinformation and harmful or illegal content. The Decision-making process must be transparent.

Civil Society

- Efforts should be made towards supporting civil society to develop digital literary and tech skills. It is important to build a civil society which understands human rights and the use of technology in the context of wider issues connected with democracy and human rights.
- By supporting civil society which will directly engage with communities impacted by disinformation, society will become more resilient to disinformation.

References:

1. ANCOM, Autoritatea Naţională pentru Administrare şi Reglementare în Comunicaţii. (2020). Decizii ANCOM pentru implementarea prevederilor Decretului nr. 195 din 16 martie 2020 şi Decretului nr. 240 din 14 aprilie 2020. *ANCOM*. https://www.ancom.ro/decizii-decret-stare-de-urgenta_6253
2. BOLETÍN OFICIAL DEL ESTADO (2020). Procedimiento de actuación contra la desinformación. https://boe.es/boe/dias/2020/11/05/pdfs/BOE-A-2020-13663.pdf
3. Bychawska-Siniarska, D. (2017). Protecting the right to freedom of expression under the European Convention on Human Rights. *Council of Europe*. https://rm.coe.int/handbook-freedom-of-expression-eng/1680732814
4. Centro Criptologico Nacional. (2019). Desinformación en el ciberespacio. *CCN-CERT*. https://www.dsn.gob.es/sites/dsn/files/CCNCERT_BP_13_Desinformaci%C3%B3n%20en%20el%20Ciberespacio.pdf
5. Charter of Fundamental Rights of the European Union (2000/C 364/01).
6. Constitution of Malta. https://legislation.mt/eli/const/eng/pdf
7. Constitution of Romania. http://www.cdep.ro/pls/dic/site2015.page?den=act2_1&par1=2#t2c2s0sba30
8. Council of the European Union. (2014). EU Human Rights Guidelines on Freedom of Expression Online and Offline. *Council of the European Union*. https://www.eeas.europa.eu/sites/default/files/eu_human_rights_guidelines_on_freedom_of_expression_online_and_offline_en.pdf
9. European Commission (2017). Communication on Tackling Illegal Content Online - Towards an enhanced responsibility of online platforms. *European Commission*. https://digital-strategy.ec.europa.eu/en/library/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms
10. European Commission (2017). Communication on Tackling Illegal Content Online - Towards an enhanced responsibility of online platforms. *European Commission*. https://digital-strategy.ec.europa.eu/en/library/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms
11. European Commission (2018). Action Plan against Disinformation. *European Commission.*

12. European Commission (2018b). COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling Online Disinformation: A European Approach. *European Commission.*

13. European Commission (2021). European Commission Guidance on Strengthening the Code of Practice on Disinformation.

14. Farrugia, D. (2020). Hybrid threats and disinformation: the COVID-19 Pandemic. https://foreign.gov.mt/en/perspectives-on-the-work-of-the-ministry/pages/hybrid-threats-and-disinformation-the-covid-19-pandemic.aspx

15. Fathaigh, R., et al. (2021). The perils of legally defining disinformation. *Internet Policy Review*. https://policyreview.info/articles/analysis/perils-legally-defining-disinformation

16. Malta's Criminal Code. https://legislation.mt/eli/cap/9/eng

17. Official Journal of the European Union. (2022). 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)

18. Romanian Law on National Security (51/1991). https://www.sri.ro/fisiere/legislation/Law_national-security.pdf

19. Romanian Penal Code. https://lege5.ro/gratuit/gezdmnrzgi/art-404-comunicarea-de-informatii-false-codul-penal?dp=gqytsojwge3te

20. Romanian Presidential Administration (2020). National Defence Strategy. *Presidential Administration*. https://www.presidency.ro/files/userfiles/National_Defence_Strategy_2020_2024.pdf

21. Romanian Presidential Administration. (2020). DECRET nr. 195 din 16 martie 2020. https://legislatie.just.ro/Public/DetaliiDocumentAfis/223831

22. Spanish Government. (2019). National Cybersecurity Strategy. *Department of Homeland Security*.

23. The European Convention on Human Rights

24. The Spanish Constitution. https://www.lamoncloa.gob.es/documents/constitucion_inglescorregido.pdf

25. Torres, M. J., Martinez-Amanasa, A., et al. (2021). Infodemic and Fake News in Spain during the COVID-19 Pandemic. *Int J Environ Res Public Health.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7918895/

26. Trynor, M. (2022). Digital Services Act (DSA) gets the green light. https://www.lewissilkin.com/en/insights/digital-services-act-dsa-gets-the-green-light-is-your-business-ready

4.2 Role of data protection in countering disinformation

Ketan Modh

Abstract

This section provides a short overview on the role of data protection in combating disinformation. In the first part, the section provides a short introduction into the right to privacy and the data protection framework. It then proceeds to examine different dimensions of data protection, which may be relevant for the prevention of disinformation, such as profiling, automated decision making, the principles of privacy by design/privacy by default, sensitive personal data. The section concludes with a short discussion on the limitations of the personal data framework in preventing and countering disinformation. The scope of the section is to provide an overview of the ways in which European data protection legislation can be employed in countering disinformation, but also the existing limitations in this field.

Main research questions addressed

- How does data protection assist in tackling disinformation?
- What are the data protection guidelines available when tackling disinformation?

The role of data protection in combating disinformation

Several techniques exist to tackle disinformation, such as content moderation and takedowns, algorithmic de-ranking, fact-checking, and media literacy. These techniques have been promoted by the European Union through its communication on tackling online disinformation (Tackling online disinformation 2018) and are crucial to combatting mass disinformation campaigns. The previous section has noted the frameworks present in the European Union and specifically in Spain, Malta and Romania that give a legal basis for implementing these techniques.[1] [2] However, a subset of problems within the fight against disinformation is that of targeted disinformation efforts – where individuals or groups of individuals are targeted with customised content based on their profiles to maximise the effectiveness and spread of such content.

To understand targeted disinformation, it is essential to place fake news within the proper context – which is that it is mainly spread via social media channels, where algorithms govern how those fake news posts are shown to users. This means that with the correct information, disinformation and fake news can spread through precisely the kinds of people who are most likely to believe it. For example, if an election management agency can gather information through social media to identify people who are more likely to support causes espoused by one political party, the agency can then target those people with fake news denouncing actions taken by the party in opposition to the one that hired it. In this manner, the agency can get people on the fence to start supporting a particular political party or encourage passive supporters to exercise their right to vote. This example is not theoretical and is precisely what Cambridge Analytica was accused of doing a few years ago (The Cambridge Analytica Files 2018).

The problem of targeted disinformation campaigns is a major one. A study on anti-vaccination disinformation in 2018 found that when Facebook altered its policies to stem the sharing of ads containing links to known fake news sites, it led to a 75% decline in the number of shares on the social media site (Chiou & Tucker, 2018). An analysis of the chain through which disinformation spreads shows that while fake content does not originate on social media websites, it is primarily through social media websites that such content is customised towards specific targets (Combatting 2021). Data protection laws help combat this aspect of disinformation. The European Data Protection Supervisor (EDPS) has also made this argument through its Opinion 3/2018 and the European Data Protection Board (EDPB) through its Guideline 8/2020. The following sections will delve into the specifics of the data protection framework and how it can be used to combat disinformation.

The Right to Privacy and Disinformation

In Europe, the right to private and family life has been set out under Article 8 of the European Convention on Human Rights (ECHR), which states as follows:

1. Everyone has the right to respect for his private and family life, his home and his correspondence.
2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

This right to private and family life in the ECHR serves to underpin the right to the protection of personal data under the EU Charter of Fundamental Rights (ECFR) through Article 8, which states that “Everyone has the right to the protection of personal data concerning him or her” (ECHR, Article 8). The Convention and the Charter have together set the stage for the EU’s data protection framework, discussed further below. In its recital, this fact has also been mentioned explicitly by the General Data Protection Regulation (GDPR), which notes that the right to personal data protection is not absolute, since it must be balanced against several other rights under the Charter. These other rights under the Charter include the respect for private and family life, home and communications (Article 7), freedom of thought, conscience and religion (Article 10), freedom of expression and information (Article 11), freedom to conduct a business (Article 16), the right to an effective remedy and to a fair trial (Article 47), and cultural, religious and linguistic diversity (Article 22). The necessity of achieving this balance arises because, as the GDPR states, all processing of personal data “should be designed to serve mankind” (GDPR, Recitals, para (4)). This data protection framework and its implications on targeted disinformation campaigns is described further below.

Data Protection Framework

In the European Union (EU), data protection falls within the aegis of the GDPR and the Law Enforcement Directive (LED), both of which were adopted in 2016 and took effect in 2018. Neither of these pieces of legislation is explicitly meant to tackle disinformation but instead broadly define the principles underlying data protection in the European Union. Of specific relevance to targeted disinformation is the 2002 Directive on privacy and electronic communications (ePrivacy Directive), which deals with the responsibilities of electronic communications services and the privacy of electronic communications.

The principles that are relevant for combating disinformation relate to profiling and automated decision making and will be dealt with separately below.

Scope of the law

The GDPR and the LED define the parties to the data processing as “controllers” and “processors” who are established in the EU or those established outside the EU but provide services within the EU. The LED specifically deals with public authorities that are competent to prevent, investigate, detect or prosecute criminal offences. However, the GDPR broadly deals with anybody (excluding the competent authorities defined in the LED) that determines the purposes and means of processing personal data. Despite dealing with separate entities, how the GDPR and LED regulate the collection and processing of personal data is essentially the same. With both the GDPR (Article 4) and the LED (Article 3), “personal data” and the “processing” of personal data are defined as follows:

1. ‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
2. ‘processing’ means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;

On the other hand, the ePrivacy Directive deals specifically with the processing of personal data in the electronic communication sector, where “communication” is defined as follows under Article 2:

> (d) “communication” means any information exchanged or conveyed between a finite number of parties by means of a publicly available electronic communications service. This does not include any information conveyed as part of a broadcasting service to the public over an electronic communications network except to the extent that the information can be related to the identifiable subscriber or user receiving the information;

The ePrivacy Directive thus complements the GDPR and is limited to a 'user' of an electronic communications service and only defines the responsibilities of such services.

In this way, the GDPR oversees the processing of a user's personal data in general, while the ePrivacy Directive oversees data processing on calls, messaging apps and services (such as via telephony services, or WhatsApp or Instagram direct messages). In each of those cases, the service providers (or "controllers") have access to the kind of data that would make it very easy to target users for advertising or other content – which could be misused for the targeted delivery of fake news. The tool used to find targets for content, that is, profiling, is discussed further below.

Profiling

The erstwhile Article 29 Working Party (Art29WP), now the EDPB, noted in its opinion on online behavioural advertising that there are two main approaches to building user profiles which can also be combined:

> i) Predictive profiles are established by inference from observing individual and collective user behaviour over time, particularly by monitoring visited pages and ads viewed or clicked on. ii) Explicit profiles are created from personal data that data subjects themselves provide to a web service, such as by registering. (Article 29, 2010)

Using such profiles for the targeted delivery of content will result in the processing of personal data as defined under the GDPR and LED. This is because profiles, whether predictive or explicit, would contain information relating to an identified or identifiable natural person. This is important because the GDPR provides explicit legal bases and principles that must be adhered to when processing personal data. Such principles, identified under Article 5 of the GDPR, include: (a) lawfulness, fairness and transparency; (b) purpose limitation; (c) data minimisation; (d) accuracy; (e) storage limitation and (f) integrity and confidentiality. Of these principles, the principle of purpose limitation is especially important and states that personal data shall be:

> (b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation'); (Article 5(1), GDPR)

The interpretation of "legitimate purposes" must be read with Article 6, which provides six bases for lawful processing of personal data. Of these six bases, two are of particular relevance in terms of targeted content delivery, and note that processing shall be lawful only if:

> (a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
>
> (f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child. (Article 6(1), GDPR)

When read in this manner, the purpose limitation principle shows that unless the user (or data subject), who is being targeted for content delivery, has provided their consent, or unless it is for pursuing legitimate interests, the processing of personal information for the targeted delivery of content would run afoul of the GDPR. The GDPR further sets out the conditions required for consent to be lawful under Article 7. In summary, the GDPR ensure that the consent is specific, informed and unambiguous, with the data subject having the right to withdraw such consent. It is important to note that even if a data subject's consent meets these criteria, the processing of personal data must still adhere to the data protection principles enshrined under Article 5 of the GDPR. This means that if the processing is not necessary and proportionate to the aimed objective, then the processing would still be invalid. This need to meet the principle of necessity and proportionality has also been noted by the Art29WP in its opinion on the conditions of consent under the GDPR (Article 29, 2010).

Through these provisions, the GDPR ensures that the data subject controls their personal data. These provisions imply that if an individual does not give their consent to be delivered targeted content, then it would contravene the GDPR to do so unless a legitimate interest (Article 6(1)(f)) can be proven. In the case of *Fashion ID*, the Court of Justice of the European Union (CJEU) provided three cumulative conditions that need to be met to prove a legitimate interest effectively:

> first, the pursuit of a legitimate interest by the data controller or by the third party or parties to whom the data are disclosed; second, the need to process personal data for the purposes of the legitimate interests pursued; and third, the condition that the fundamental rights and freedoms of the data subject whose data require protection do not take precedence (CJEU, 2019, C-40/17, para. 95)

If these three conditions are met, then the legal basis of legitimate interest under the GDPR may be used. Even so, as mentioned earlier, any processing of personal data must meet all six principles under Article 5, including the principle of necessity and proportionality. This has also been made clear by the EDPB in its opinion on the targeting of social media users(EDPB 2020).

In terms of predictive profiles, which are based on observing individual and collective user behaviour over time, it is necessary to look at the ePrivacy Directive, which deals specifically with tracking data in electronic communications. This is because, in general, the observation of user behaviour occurs through tracking technologies such as cookies which are stored on the user's device. Article 5(3) of the ePrivacy Directive states that:

Member States shall ensure that the use of electronic communications networks to store information or to gain access to information stored in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or user concerned is provided with clear and comprehensive information in accordance with Directive 95/46/EC, inter alia about the purposes of the processing, and is offered the right to refuse such processing by the data controller.

Directive 95/46/EC noted in the language above was superseded by the GDPR, and therefore all references to the Directive now point to the GDPR instead. The EDPB has given its opinion on the interplay between the ePrivacy Directive and the GDPR, noting that the interpretation of the ePrivacy Directive must be along the narrow lines drawn by the GDPR (EDPB, 2019). Under Article 5(3) of the ePrivacy Directive, the user must be provided "clear and comprehensive" information on the purposes of processing personal data. This is similar to the principle of transparency under Article 5 of the GDPR. Further, the processing of personal data under the ePrivacy Directive is based on the user's consent or, in the case of the GDPR, the data subject's consent. The conditions of such consent must also be interpreted following Article 7 of the GDPR discussed above, meaning that the consent must be specific, informed and unambiguous, with the data subject having the right to withdraw such consent.

Taking this into account, whether the processing of personal data is through predictive profiles or explicit profiles, adhering to the principles of data protection set out in the GDPR is critical. Having said that, there are further limitations on decisions taken through the use of these profiles, which are discussed below.

Automated Decision Making

As discussed earlier, for targeted disinformation campaigns, user profiles are used to decide how to customise content and whom to deliver that content. The GDPR has specific restrictions on making such automated decisions under Article 22 which states:

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:

> (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>
> (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>
> (c) is based on the data subject's explicit consent.

For the purposes of targeted content delivery, Article 22(2)(c) becomes relevant due to the need for explicit consent. In its opinion on the conditions of consent, the Art29WP defines it as such:

> The term explicit refers to the way consent is expressed by the data subject. It means that the data subject must give an express statement of consent. An obvious way to make sure consent is explicit would be to expressly confirm consent in a written statement (Article 29, 2016, 18)

This is different from 'regular' consent because regular consent only requires a statement or clear affirmative action per the GDPR. Thus, to be subjected to automated decision-making or profiling, a data subject must provide explicit consent. If such explicit consent is not obtained by the controller, they would be in contravention of the GDPR and would therefore be unable to delivery targeted content to the data subject.

The need to record consent becomes especially important when dealing with sensitive categories of personal data or when designing the data collection system, as will be discussed below.

Sensitive Categories of Personal Data

Under the GDPR, some types of personal data have been designated as "special categories" under Article 9, which states:

> 1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

It is clear that "special categories" actually refer to data about a person that is highly sensitive in nature or data about a person that forms a core part of a person's identity. This is the type of data that is especially susceptible to misuse in the case of targeted delivery of content. For example, knowing that one has a particular political belief would make those persons easy targets of opportunity for those trying to sway them into voting a certain way.

Any personal data that falls under the ambit of this Article can only be processed under specific circumstances provided under Article 9(2), of which two are relevant for the purposes of targeted content delivery:

> (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject;
>
> (e) processing relates to personal data which are manifestly made public by the data subject;

In terms of Article 9(2)(a), the interpretation of explicit consent is exactly as discussed above. Therefore, unless explicit consent has been given, special categories of personal data cannot be used to delivery targeted content to data subjects.

Article 9(2)(e) is relevant because targeters of the content may state that they used special categories of personal data to target content delivery since such information had already been made public by the data subject. However, the standard for 'manifestly made public' has been discussed at length by the EDPB, which requires a combination of several elements to demonstrate that this standard has been met. These elements include the default settings of the platform where information has been published, the nature of that platform, the accessibility of the page where such data has been made available, the visibility of the information, and whether the data subject himself or herself has published the sensitive data (EDPB, 2020, 35).

Data Protection by Default and by Design

Taking into account all the principles and legal bases described above, possibly the most critical safeguard for data subjects from an implementation perspective has been set out under Article 25. This article states as follows:

> 1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the

> processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.
>
> 2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

Especially in terms of the targeted delivery of content, the need to ensure data protection by default means that data subjects would need to provide their consent (whether regular or explicit) as an ‘opt-in’ measure; that is, the consent would not be provided by default. This also means that service providers would be obligated to ensure that a user would have to specifically choose to divulge their personal information instead of giving away that data as the default option. Further, their data would be protected by any safeguards deemed necessary by the GDPR through the design of the data collection system itself. Thus, only the minimal amount of data necessary and proportionate to meet a controller’s objective would be collected; this data would only be used for a limited purpose and then deleted after a specific amount of time. As long as these measures are in place, the targeted delivery of content would only take place for those data subjects who opt-in to it or where the controller can demonstrate a legitimate interest in doing so.

Impact of Data Protection on Disinformation

Drawing on the discussion of the relevant data protection principles above, it is necessary to discuss them in the context of disinformation. This leads to some important conclusions regarding their limitations. First, most websites and services that attract a large user base have created entire businesses out of delivering targeted content. Facebook and Google (Constine, 2018), two of the most extensive internet services on the planet, have already demonstrated an acute desire to soften the impact of the data protection principles under the GDPR, for which they have also been fined. This is despite, for example, Facebook being at the heart of the Cambridge Analytica scandal, which revolved around targeted disinformation campaigns based on profiling users (The Cambridge Analytica Files 2018).

Second, while data protection guidelines can provide a measure of protection to data subjects and safeguard their right to privacy, these guidelines also depend on data subjects exercising sensible control over their personal data. This can be seen through data protection hinging on the need for regular or explicit consent in the case of profiling and automated decision-making. Unfortunately, the methods employed by websites to comply with the conditions of consent have led to “consent fatigue”,[1] where users have simply stopped paying attention to consent notifications. Combined with internet services trying their best to get user consent, this has led to a situation where the usual means of obtaining consent are being called into question by bodies such as the Belgian Data Protection Authority, which recently issued a €250,000 fine to a company providing consent-related services[2].

While acknowledging the limitations of data protection guidelines, it is necessary to note that a large amount of disinformation occurs through the targeted delivery of content. For example, echo chambers (Cinelli et al, 2021) have been identified as one of the most effective methods of spreading disinformation (Menczer, 2016). These echo chambers can be taken advantage of only by processing personal data, thus allowing data protection guidelines to mount a practical challenge to this disinformation vector. When filter bubbles and echo chambers in social media have been identified as a vital part of the fake news and disinformation ecosystem (Rhodes, 2022) data protection guidelines become all the more important in combating the phenomenon, regardless of the limitations of such guidelines as discussed above. Therefore, it is essential to follow the guidelines, especially that of data protection by design and by default, to ensure that the limitations are covered, and disinformation is effectively tackled.

References:

1. Article 29 Working Party available at Article 29 Working Party Guidelines (dataprotection.ro)
2. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling online disinformation: a European Approach COM/2018/236 final
3. Chiou, L., & Tucker, C. (2018). *Fake news and advertising on social media: A study of the anti-vaccination movement* (No. w25223). National Bureau of Economic Research.
4. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118.
5. Combatting Targeted Disinformation Campaigns (dhs.gov) 2021
6. Constine, J. (2018) A flaw-by-flaw guide to Facebook's new GDPR privacy changes, available here: https://techcrunch.com/2018/04/17/facebook-gdpr-changes/
7. Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA
8. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)
9. European Data Protection Supervisor, EDPS Opinion on online manipulation and personal data, available here: 18-03-19_online_manipulation_en.pdf (europa.eu)
10. European Data Protection Board, Guidelines 8/2020 on the targeting of social media users, available here: EDPB guidelines: cookies, consent and compliance (cookiebot.com)
11. European Convention on Human Rights (ECHR), available at European Convention on Human Rights (coe.int)
12. European Data Protection Board, Guidelines 8/2020 on the targeting of social media users, available here: https://edpb.europa.eu/system/files/2021-04/edpb_guidelines_082020_on_the_targeting_of_social_media_users_en.pdf
13. EDPB Opinion 5/2019 on the interplay between the ePrivacy Directive and the GDPR available at Opinion 5/2019 on the interplay between the ePrivacy Directive and the GDPR, in particular regarding the competence, tasks and powers of data protection authorities | European Data Protection Board (europa.eu)

14. JUDGMENT OF THE COURT (Second Chamber) 29 July 2019 available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62017CJ0040&from=EN
15. Menczer, F. (2016). Fake online news spreads through social echo chambers. *Scientific American*, *28*.
16. REGULATION (EU) 2016/ 679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation) (europa.eu)
17. Rhodes, S. C. (2022). Filter bubbles, echo chambers, and fake news: how social media conditions individuals to be less critical of political misinformation. *Political Communication*, *39*(1), 1-22.
18. The Guardian, *The Cambridge Analytica Files*, 2018 available at: https://www.theguardian.com/news/series/cambridge-analytica-files

4.3 Case Studies

Kethan Modh, Aitana Radu, Valentin Stoian-Iordache, Cristina Arribas, Manuel Gertrudix, Ruben Arcos

Abstract

This analysis looks at relevant national and European case law, where available, to determine how courts have interpreted (or, in some cases, made up for the lacunae in) the legislation in the field of fake news and disinformation in three countries. This is especially important since there has been a sharp rise in disinformation and fake news since the COVID-19 pandemic, especially with conspiracy theories regarding vaccination (otherwise called the ‘infodemic’). The aim of this section is to carry out a review of existing national (Malta, Romania and Spain) and European case law, where available, and present some of the challenges and best practices encountered by the judicial system in addressing cases of fake news dissemination and/or disinformation.

Main research questions addressed

- What major disinformation-related cases have occurred in Malta, Romania and Spain?
- How have Malta, Romania and Spain addressed these disinformation-related cases?

General Principles Derived from Case law

While previous sections of this chapter have dealt with the legislative tools utilised in Malta, Romania and Spain to tackle disinformation, the current section looks at relevant case law, where available, to determine how courts have interpreted (or, in some cases, made up for the lacunae in) the legislation in each country. This is essential since there has been a sharp rise in disinformation and fake news since the COVID-19 pandemic, especially with conspiracy theories regarding vaccination (otherwise called the ‘infodemic’).

Given the recent nature of the infodemic, as well as the relatively modern phenomenon of the quick spread of false information through social media, there is a severe dearth of case law in the European Union directly on disinformation, including in the European Court of Human rights (ECtHR) and in the Court of Justice of the European Union (CJEU). A few characteristics of the jurisprudence surrounding online disinformation and fake news must be highlighted.

Firstly, there is a wealth of relevant case law that indirectly deals with disinformation, such as lies and falsities (created by state and private actors) that result in a violation of human rights. This is especially true with case law related to journalists and editors regarding the burden of responsibility they bear in ensuring the accuracy of published information. However, case law dealing with individuals spreading lies has only really been tackled in the realm of defamation. Indeed, the first mention of ‘fake news’ indeed being introduced by the court without any prompting from the parties involved in the case of *Brzeziński v. Poland* (ECtHR App. No. 47542/07, 2019). *[3]*. In this case, an election booklet criticising and/or defaming government members published by a local politician was characterised as false and defamatory information and ruled as such by local courts. When examined by the ECtHR, the court looked at the issue in terms of the right to freedom of expression under Article 10 of the European Convention on Human Rights (ECHR). The ECtHR noted that while it was undoubtedly necessary to ensure that a forthcoming election not be influenced by ‘fake news’, the way the local courts went about doing so had a chilling effect on the applicant’s right to the freedom of expression.

Secondly, the spread of fake news and disinformation on the internet has generally been dealt with in light of the specific features of the internet, i.e., its amplifying effect. For example, in *Cicad v Switzerland[4]*, a case where an allegation of antisemitism was made by an association on its website, the association was ordered to remove the content since it would be visible to an audience far more expansive than its usual offline audience (ECtHR App. No. 17676/09, 2016). The nature of the allegation being made online rather than offline was a deciding factor in the Court asking for the content to be removed.

Thirdly, the nature of the ‘speech’ being made is also relevant. One of the vectors used in the spread of fake news is through users of social media highlighting a post for other users by ‘reacting’ to it, such as by ‘liking’ it or sharing it. In the case of *Melike v Turkey[5]*, the Court considered a person’s use of pressing the ‘Like’ button on social media vis-à-vis sharing a post, and also took into account the size of the audience that takes note of a user pressing the ‘Like’ button (ECtHR App. No. 35786/19, 2021). In this case, the applicant’s use of the button on a post containing virulent political criticism was seen in terms of the penalty imposed (she was fired from her post), with the penalty being disproportionate and thus violative of her right to freedom of expression under Article 10 of the ECHR.

Keeping in mind the nature of jurisprudence on fake news and disinformation, this chapter will now examine important cases in Spain, Malta, and Romania.

4.3.1 Case study in Spain

Spain is an important country to consider when dealing with disinformation, specifically with the removal of online content as a tool to fight disinformation. It was through the case of *Google Spain SL v. Agencia Española de Protección de Datos[6]*, otherwise known as *Google v Spain*, that the “derecho al olvido” (right to be forgotten) was first devised by the CJEU. The issue at the heart of this case was the removal of content that was no longer accurate (CJEU (Grand Chamber), Case C-132/12 ECLI:EU:C:2014:317). The passing of this judgment ensured that the right to be forgotten was incorporated in the General Data Protection Regulation (GDPR), which was only being drafted when the judgment was passed. In particular, the Court held that:

> It follows from those requirements, laid down in Article 6(1)(c) to (e) of Directive 95/46, that even initially lawful processing of accurate data may, in the course of time, become incompatible with the directive where those data are no longer necessary in the light of the

> purposes for which they were collected or processed. That is so in particular where they appear to be inadequate, irrelevant or no longer relevant, or excessive in relation to those purposes and in the light of the time that has elapsed (CJEU (Grand Chamber), Case C-132/12 ECLI:EU:C:2014:317, para 93).

Further, the court laid down the context surrounding the right to be forgotten in the following terms:

> In the light of the foregoing, when appraising such requests made in order to oppose processing such as that at issue in the main proceedings, it should in particular be examined whether the data subject has a right that the information relating to him personally should, at this point in time, no longer be linked to his name by a list of results displayed following a search made on the basis of his name. In this connection, it must be pointed out that it is not necessary in order to find such a right that the inclusion of the information in question in the list of results causes prejudice to the data subject ((CJEU (Grand Chamber), Case C-132/12 ECLI:EU:C:2014:317, para 96)

This right now exists in the GDPR under Article 17 (Right to erasure (‘right to be forgotten’)) and Article 19 (Notification obligation regarding rectification or erasure of personal data or restriction of processing). This is one of the ways that disinformation regarding specific individuals can be fought – by invoking the right to erasure in case of false information.

The other aspect of the fight against disinformation, that of ‘spreading lies’, has recently seen some movement in Spain. In early November 2022, a member of the Guardia Civil, a law enforcement agency, was sentenced to 15 months in prison, along with a fine, by a local court in Barcelona for falsely claiming that a video he linked to was showing an underage migrant trying to rape a woman (Jones, 2022). The original video was from China, but was viewed over 22,000 times along with the law enforcement officer’s false description.

Disinformation has become a real issue in Spain; a recent survey on the proliferation of false news on Covid-19 concluded that “the vast majority of those surveyed attributed a high level of repercussion to the hoaxes related to COVID-19, considering the social impact generated by the resulting alarm situation to be serious or very serious” (Torres et al., 2021). On the other hand, there has been concern shown about the use of the Penal Code to criminalise jokes as well, with calls by civil society organisations for law enforcement to ‘refrain from using criminal prosecution and other coercive measures as the primary means of combating supposedly false or harmful information online’ (Article 19, 2020).

4.3.2 Case study in Malta

While the country has not been affected by disinformation campaigns during Covid-19 to the extent of others in Europe, Malta does not have a good track record with protecting media freedom, let alone tackling fake news and disinformation. A Venice Commission report in 2018[12] noted (Venice Commission, 2018):

> The media and civil society are essential for democracy in any state. Their role as watchdogs is an indispensable precondition for the accountability of Government. The delegation of the Venice Commission had the impression that in Malta the media and civil

> society have difficulty in living up to these needs (Venice Commission, 2018, para 135).[13]

This is important to note given the recent disinformation campaigns levelled against journalists in Malta, as noted by the International Federation of Journalists (International Federation of Journalists, 2021).[14]. Malta has instead taken the tack of amending the provision criminalising the spreading of false information under Art. 82 of the Criminal Code by expanding it through the Media and Defamation Act 2018.[15] Initially, this article only criminalised the act of maliciously spreading false news, but through the amendment in 2018, Malta has also levied penalties of imprisonment and a fine if this offence "has contributed to the occurrence of any disturbance", clearly planning to include the spread of fake news and disinformation.

This was likely a reaction to the death of Daphne Caruana Galizia, a Maltese journalist, near the end of 2017. Indeed, this event has been the focus of several disinformation campaigns (see section 1.3), due to both the political nature of the subject matter of the journalist's reportage and the shocking nature of her death. Having said that, no judgments have been passed in Malta on the basis of Art. 82 of the Criminal Code, though several complaints have been filed by the police, including one recently by the President of Malta (Malta Independent, 2022).[16]. How Malta proceeds with such complaints from a judicial perspective is yet to be seen.

4.3.3. Case Study in Romania

Disinformation has played a significant role in Romania during Covid-19, specifically concerning vaccine hesitancy. This hesitancy has been identified as one of the key reasons for the fourth wave of the pandemic in Romania, both by news (Euronews, 2021)[17] and scholarly[18] sources (Dascalu et al., 2021). As one source notes:

> […] throughout the pandemic, under the pretext of presenting "balanced viewpoints", major news outlets generously featured representatives of the anti-vaccine movement and conspiracy theory advocates almost on a daily basis(Dascalu et al., 2021, para 2).[19]

The state of emergency declared in Romania on 16 March 2020 allowed the government to take down instances of fake news (a policy that the government implemented very proactively[20])(Euractiv, 2020); this was clearly not very effective in curbing anti-vaccination conspiracies. This was because major sources of this disinformation included religious leaders and politicians, with one Bishop being placed under criminal investigation for spreading disinformation based on his anti-vaccination comments (Bdnews2.com, 2021) [21]. As the article notes, this disinformation led to Romania's vaccination rate being under 30% in late 2021 (vis-à-vis the European average of 81%), resulting in a devastating fourth wave of Covid-19 with the highest per-capita death rate in the world.

In conclusion, one of the major problems with using case law in determining the effectiveness of legislative tools is that cases of disinformation or fake news are tough to prosecute for several reasons. Firstly, laws that combat disinformation via the route of defamation rely on having someone to prosecute or assign blame to for spreading false information. This is difficult simply given the nature of the internet. To take the example of Malta, there was clearly a disinformation campaign against journalists who reported on the consequences of Daphne Caruana Galizia's death. However, this disinformation campaign was perpetrated through fake accounts and websites, making it difficult to pin the blame on any individual.

Secondly, given that these laws fall under the criminal code of most countries, it also becomes necessary to show that the person spreading the fake news did so 'maliciously' or with the intent to cause harm. This is difficult to determine even when done on a case-by-case basis. For example, in Romania, would it be possible to determine whether the religious leaders or politicians wanted to cause harm by protesting against vaccinations?

Thirdly, ECtHR case law makes it clear that restricting someone's right to the freedom of speech can only be done keeping the specific context of the situation in mind, which means that a blanket ban (or internet shutdown) is, quite rightly, illegal. However, given the scope of the problem, it would be impossible to prosecute every a single person who "Liked" or shared a post that happens to contain false information without bogging down the entire judicial system.

Given the fact that disinformation and fake news are rapidly becoming the primary factors in undermining democratic processes and human rights[22], it is incumbent upon the judiciary, in tandem with the legislature, to determine the way forward despite these challenges. We will likely observe significant progress here in the near future simply because of the amount of effort dedicated to solving the issue of disinformation worldwide, but this will be a daunting process (European Parliament, 2021).

References:

1. Article 19. (2020). Spain: Concerns as Penal Code used to criminalise jokes and misinformation about coronavirus. *Article 19.* https://www.article19.org/resources/spain-penal-code-used-to-criminalise-jokes-and-misinformation-about-coronavirus/
2. Bdnews2.com.(2021). In Romania, hard-hit by COVID, doctors fight vaccine refusal. *Bdnews2.com.* https://bdnews24.com/world/europe/in-romania-hard-hit-by-covid-doctors-fight-vaccine-refusal
3. CJEU (Grand Chamber). (2014). Case C-132/12 ECLI:EU:C:2014:317.
4. Dascalu, S., Geambasu, O., Valentin Raiu C., Azoicai, D., et al. (2021). COVID-19 in Romania: What Went Wrong?. *Front. Public Health.* doi: 10.3389/fpubh.2021.813941
5. Euractiv. (2020). Romania shuts down websites with fake COVID-19 news. *Euractiv.* https://www.euractiv.com/section/all/short_news/romania-shuts-down-websites-with-fake-covid-19-news/
6. Euronews (2021). Why did Romania's vaccination campaign derail after such a good start?. *Euronews.* https://www.euronews.com/my-europe/2021/06/08/why-did-romania-s-vaccination-campaign-derail-after-a-successful-start
7. European Court of Human Rights. (2016). App. No. 17676/09.
8. European Court of Human Rights. (2019). App. No. 47542/07.
9. European Court of Human Rights. (2021). App. No. 35786/19.
10. International Federation of Journalists. (2020). Malta: Journalists and public figures harassed in disinformation campaign. *International Federation of Journalists.* https://www.ifj.org/media-centre/news/detail/category/press-releases/article/malta-journalists-and-public-figures-harassed-in-disinformation-campaign.html
11. Jones, S. (2022). Spanish police officer sentenced after posting fake rape video on Twitter. *The Guardian.* https://www.theguardian.com/world/2022/nov/08/spanish-police-officer-sentenced-after-posting-fake-video-on-twitter

12. Malta Independent. (2022). President’s office files police report on fake article. *Malta Independent.* https://www.independent.com.mt/articles/2022-07-28/local-news/President-s-office-files-police-report-on-fake-article-6736244815

13. Torres, M. J., Martinez-Amanasa, A., et al. (2021). Infodemic and Fake News in Spain during the COVID-19 *Pandemic. Int J Environ Res Public Health*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7918895/

14. Venice Commission. (2018). Malta - Opinion on Constitutional arrangements and separation of powers and the independence of the judiciary and law enforcement, adopted by the Venice Commission at its 117th Plenary Session - CDL-AD(2018)028-e.

5. TECH-DRIVEN SOLUTIONS AND EMERGING TECHNOLOGIES TO COUNTER DISINFORMATION

Introduction

This chapter aims at presenting the main technological instruments and initiatives in the field of combating online disinformation, playing a significant role in the general architecture of the handbook, helping the target group understand the phenomenon of disinformation in a comprehensive manner, by acknowledging which are the weapons one can use in order to avoid becoming a victim of false content. Therefore, the chapter will try to display the main features of different technical solutions to combat online disinformation, also aims at developing the readers' digital skills by exploring.

Therefore, this chapter will focus on (1) identifying and briefly presenting the main directions that can be followed in order to combat the negative effects of disinformation through the use of technology; (2) defining the main concepts used within this chapter, so as to establish a common theoretical background for the approached directions; (3) briefly presenting the main technological solutions available on the market for both raising the level of awareness of the public opinion (at large) towards the effects of online disinformation and digitally training the public opinion to avoid becoming a victim of the online disinformation phenomenon; and (4) examining the limitations that the technological solutions have when it comes to detecting disinformation. Therefore, the chapter will define concepts such as media literacy, serious games etc., presenting their applicability in the field of disinformation. This section will also briefly analyse core technological tools used for detecting and countering disinformation and discuss their potential and limitations, providing a brief analysis on the human rights impact of such tools.

The result of this chapter can be structured into three main areas: (1) a first part dedicated to establishing a common understanding of the key concepts that will be developed and employed in the whole chapter; (2) a compiled list of one of the most known applications and other technological solutions to counteract and combat (to a certain level) the effects of online disinformation; (3) the main limitations and challenges posed by technological solutions to countering online disinformation.

Digital competencies addressed:

1.1 Browsing, searching and filtering data, information and digital content;
1.2 Evaluating data, information and digital content;
1.3 Managing data, information and digital content;
2.1 Interacting through digital technologies;
2.2 Sharing through digital technologies;
2.3 Engaging citizenship through digital technologies;
2.6 Managing digital identity.

5.1. Combating the effects of disinformation in the online environment

Alexandra Anghel, Ana Ćuća, Aitana Radu

Abstract

Recently, both technological developments and advances registered in the sector of Internet of Things and social networks have created the premises for the expansion of the noxious effects of the disinformation phenomena. Even though disinformation is not considered a recent phenomenon, it gained widespread attention from governments worldwide as a result of its applicability in the online environment, throughout the use of social media. Therefore, topics such as *fake news*, *disinformation*, *propaganda* have experienced a resurgence of interest in nowadays societies, resulting from the generalized concerns around the widespread negative effects and impact of disinformation on public opinion and public events (Sharma, et al, 2019, 2).

Furthermore, the rise of ubiquitous misinformation, disinformation, deepfakes, and post-truth raised also increasing concerns on the role of the Internet and social media in current democratic societies. Given its fast and widespread diffusion, disinformation imposes not only an individual and societal cost, but can also determine economic losses or national security risks (Fraga-Lamas & Fernandez-Carames, 2020, 53). In addition, the features of online communication, such as the speed and scope at which false information can be disseminated in the online environment, led to an increased potential of the people to deceive throughout the usage of computer-mediated communication channels, aspect that can produce major changes on financial markets, as well as political scenes (Fuller, Biros, & Wilson, 2009).

Taking into account the above-mentioned aspects, it is true to say that technology has created the means for the expansion of the disinformation phenomenon, social media becoming one of the main sources of information for the population at large, as well as an important source of false content and digital deception. However, technology can also play an essential role in combating the effects of online disinformation and propaganda and in containing the expansion processes of these now defined security issues. This chapter aims, therefore, at presenting the main technological instruments and initiatives in the field of combating online disinformation, in addition to the previous sections that present the techniques employed by different disinformation processes. This chapter plays a significant role in the general architecture of the handbook, helping the target group understand the phenomenon of disinformation in a comprehensive manner, by acknowledging which are the weapons one can use in order to avoid becoming a victim of false content, as well as developing their digital skills by exploring different technical solutions to combat online disinformation.

This section is structured into two main parts: (1) a part dedicated to establishing a common understanding of the key concepts that will be developed and employed in the whole chapter, and (2) a compiled list of one of the most known applications and other technological solutions to counteract and combat (to a certain level) the effects of online disinformation.

Main research questions addressed

- What is the role of technological tools in fighting disinformation?
- Which are the main technological initiatives in in terms of combating disinformation and what is their potential and limitations?

Combating the effects of disinformation in the online environment – setting the scene

Since 2016, the world is actively witnessing the rise of disinformation across different media platforms. The reason behind it lies in the intersection of several factors. Firstly, since 2016, there is increasing online propaganda disseminated through hyper partisan news sites that use disinformation as a business model for generating profit. During the 2016 US elections, the number of “news sites" that fabricated pro-Trump news skyrocketed. Moreover, these news sites were not restricted to the US, but could even be found in remote parts of the world. For example, over 100 pro-Trump websites were registered in the small town of Veles in North Macedonia, which produced viral fake stories promoting Trump (Posetti, 2018). In 2017, Facebook confirmed that Russia spent over $100,000 to finance ads which spread polarizing views on different societal topics, such as immigration, race and LGBTQI rights, all of which were topics of discussion during the 2016 presidential elections (Menn & Ingram, 2017). Secondly, politicians are increasingly using propaganda terms to frame political issues, instead of employing a fact-based approach. In 2017 alone, former US president Donald Trump made over 1,999 false or misleading claims (Kessler et al., 2019). Every time Trump, or any other politician repeats misleading claims, even when these have been proven to be untrue by the media, they still have a significant impact on public trust. Thirdly, the technological advancement in the field of advertising algorithms and social media platforms enabled the creation of partisan camps and polarized crowds. Search engine optimization, personalized social media feeds, and micro-targeted advertising allow polarized crowds to consume content that confirms their prior beliefs (Kessler et al., 2019).

Moreover, there is no significant difference between exposure to fake news in Europe vs. the US. According to the results of a 2018 public consultation organized by the European Commission, out of 2986 participants, 97% of them claimed to have been exposed to disinformation, 38% on a daily basis, and 32% on a weekly basis. The majority of participants (74%) believed that social media facilitates the spread of disinformation (European Commission, 2018).

The issue of disinformation has become even more prominent in the context of the Ukrainian war. Both the previous Donbas conflict in Ukraine and the current war show how countries such as Russia are extensively engaging in information warfare via social media platforms.

One potential answer to this growing trend of online propaganda and disinformation is the use of fact-checking by the media, think tanks and/or individual experts (see also section 3.4 for more information on fact-checking). However, fact-checking can be a lengthy process that includes several stages, such as:

1. Identification: Includes constant media monitoring and constant monitoring of political sources. Given the amount of news that are published daily, this stage also includes prioritization of claims who urgently need to go through the fact-checking procedure;
2. Verification: Includes checking the identified claim against an already existing fact-check as well as checking the information against official sources. In some occasions, the verification stage will also include credibility sourcing.
3. Correction: Includes flagging the false information, providing additional data to contextualize provided information, and publishing the corrected news (Alphilippe et al., 2019).

Given the lengthiness of the fact-checking procedure and the amount of disinformation that is continuously produced and shared, there is a growing need for using automated fact-checking tools and other tech-driven solutions. Development of such tools allows interested groups (fact checkers, media, academia, policymakers) to understand the most important news/claims that need to be fact-checked, to timely react if someone is sharing disinformation and, most importantly, to detect disinformation that is starting to circulate.

Apart from fact-checking, there is a growing need for detecting social media bots which manipulate online discussions and are often used for spreading disinformation and manipulating narratives (Alphilippe et al., 2019). Whereas technology can be used to amplify disinformation on social networks either through the creation and promotion of disinformation or through the use of social media bots, tech-driven solutions are also leading the way in the fight against disinformation.

The growth of computer-mediated communication though the usage of social media registered during the last decade, as well as the changes produced in the terminology specific of the disinformation phenomena led to an evolution of the core nature and characteristics of the problem itself (Sharma, et al, 2019, p. 5). In order to better understand the way in which technology can be employed to counteract the negative effects of disinformation, it is important to acknowledge the factors that allow fake news and disinformation to spread at both individual and social level.

In consequence, in addition to the factors addressed in chapters 2, when referring to the individual level, the literature in the field has shown that the inability of a person to discern in an accurate manner false content from real one led to an uninterrupted process of sharing and believing of false information disseminated on social media (Sharma, et al, 2019, 5). As an example, a survey conducted by the international research, data and analytics group YouGov in 2017 on 1684 British adults who were required to analyze the credibility of six individual news stories (half of which were false and the other half true) found that only 4% of the individuals had the capacity to correctly identify them (Channel 4, 2017). The inability to distinguish false content from the real one was attributed to ideological biases and cognitive abilities. In addition, authors Gordon Pennycook and David G. Rand presented in their study the positive correlation between propensity for analytical thinking and the ability to differentiate false from true content (Pennycook & Rand, 2019). However, authors Hunt Allcott and Matthew Gentzkow showed that there are differences in the way in which people perceive information available on social media, generated by the time allotted for consuming media content, their level of education and their age (with higher educated, older people being more accurate in forming perceptions of information) (Allcott & Gentzkow, 2017, 228).

Adjacent to cognitive abilities, ideological priors can also play an important role in the process of information consumption. Naive realism (which refers to the individual tendency to trust more easily in information that is aligned with his/her own views), confirmation bias (individuals tend to select and prefer to receive only that information which confirms their existing views, rejecting any piece of information that contravene their points of view), and normative influence theory (individuals have the tendency to disseminate and consume socially safe options as a preference for social acceptance and affirmation) are generally considered important elements in the perception and dissemination of fake news and disinformation (Shu, Sliva, Wang, Tang, & Liu, 2017, 24).

As the social level, the core of social media and collaborative information sharing on online platforms provides a supplementary dimension to disinformation and fake news, generally known as the echo chamber effect (Shu, Sliva, Wang, Tang, & Liu, 2017, 225). The principles of naive realism, confirmation bias, and normative influence theory stated above, describe the need of individuals to search, consume, and disseminate information that is in alignment with their own viewpoints and ideologies, developing, in consequence, the tendency to establish and develop connections with ideologically similar individuals (social homophily). Therefore, social media algorithms focus on customizing recommendations (algorithmic personalization) by suggesting content that better fits an individual's preferences, as well as by recommending connections to persons that share similar beliefs (Sharma, et al, 2019, 5). Both social homophily and algorithmic personalization contribute to the development of echo chambers and filter bubbles, wherein individuals get less exposure to conflicting viewpoints and become isolated in their own information sphere (Garimella, Gionis, Parotsidis, & Tatti, 2017, 4663). The existence of echo chambers can increase the chances of survival and continuous dissemination of fake news, aspect that can be explained by the phenomena of social credibility and frequency heuristic. The concept of social credibility indicates that people's perception of credibility of a piece of information tends to increase if others also perceive it as credible (especially in those cases when the credibility of the source of information cannot be tested), and the frequency heuristic concept defines the tendency to grant a higher level of credibility to a piece of information to which an individual is exposed multiple times (Shu, Sliva, Wang, Tang, & Liu, 2017, 25).

Given these factors, as well as those referred to in chapters 1 and 2, one can conclude that technology has equipped the general public with new and more developed capabilities to consume different media contents, from streaming video to reading niche blogs and news sites. In addition, technology also developed for people worldwide the habit of trusting the transparency of the content they consume each day from social media, without questioning or evaluating the source of information (Fowler, 2022). However, technology did not only create the premises for the expansion of online disinformation, but it also allowed the development of solutions to combat the negative effects of the above mentioned phenomena.

The majority of tech-driven solutions are relying on machine learning. The attractiveness of machine learning in the context of targeting and combating disinformation arises from the fact that machine learning models can recognize novel cases and react to them, based on prior learning. The possibility of continuous improvement of machine learning models, makes them seem like an effective tool to address the always-evolving world of disinformation. In this context, the next section of the chapter will map the main assessment methods used for developing technological instruments to combat online disinformation (identified based on an analysis on the existing literature) and current efforts from the Machine Learning (ML) community to fight against the threats posed by the disinformation phenomenon at both individual and social level.

Assessment methods – a review

In order to set the framework for a better understanding of the main solutions identified in the domain of combating the negative effects of online disinformation, it is essential to define the assessment methods that the most majority of technological solutions are based on. Therefore, the literature in the domain divides the methods, which emerged from various domains, using disparate techniques, into two main categories (see Conroy, Rubin, & Chen, 2015):

(1) *Linguistic approaches* – focus on extracting and analyzing the content of deceptive messages in order to associate language patterns with deception. More specifically, this type of approach is aimed at identifying „leakages" in the content of the message analyzed by measuring the frequency and patterns of pronouns, conjunction, negative emotion word usage and so on (Feng & Hirst, 2013). The methods associated with this category are, as follows:

Method	Brief presentation
Data representation	One of the simplest methods of representing texts is the "bag of words" approach, which regards each word as a single, equally significant unit. In the bag of words approach, individual words or "n- grams" (multiword) frequencies are aggregated and analyzed to reveal cues of deception. However, by relying on isolated n-grams, often divorced from useful context information, any resolution of ambiguous word sense remains non-existent (Conroy, Rubin, & Chen, 2015, 2).
Deep syntax	Since the analysis of word use is sometimes not enough in predicting deception, deeper language structures (syntax) have been developed as a complementary solution. Deep syntax analysis is implemented through Probability Context Free Grammars (PCFG), that transforms sentences in a set of rewrite rules (a parse tree) to describe syntax structure (e.g. noun and verb phrases), which are in turn rewritten by their syntactic constituent parts (Feng, Banerjee, & Choi, 2012). The final set of rewrites produces a parse tree with a certain probability assigned. The method is used to distinguish rule categories (lexicalized, un-lexicalized, parent nodes etc.) for deception detection with 85-91% accuracy (depending on the rule category used) (Conroy, Rubin, & Chen, 2015, 2). However, used alone, this method

	might not be sufficiently capable of identifying deception, therefore studies often combine this approach with other linguistic or network analysis techniques (Feng, Banerjee, & Choi, 2012) (Feng & Hirst, 2013).
Semantic analysis	This approach extends the n-gram plus syntax model by incorporating profile compatibility features, showing the addition significantly improves classification performance (Feng & Hirst, 2013). This method uses the principle of aligning profiles and the description of the writer's personal experience, in order to assess veracity based on compatibility scores: 1. compatibility with the existence of some distinct aspect (e.g. an art museum near a mentioned hotel); 2. compatibility with the description of some general aspect, such as location or service. In this case, the prediction of falsehood is shown to be at approximately 91% accurate (Conroy, Rubin, & Chen, 2015, 2)
Rhetorical Structure and Discourse Analysis	A method used to achieve the description of discourse, by identifying the instances of rhetoric relations between linguistic elements (Rubin & Lukoianova, 2014)
Classifiers	A mathematical model sufficiently trained from pre-coded examples in a specific category, that can predict instances of future deception on the basis of numeric clustering and distances (Conroy, Rubin, & Chen, 2015, 3)

Table 8 Overview of methods employed for linguistic analysis

(2) *Network approaches* – focus on the usage of network properties and behavior to complement content-based approaches that rely on deceptive language and leakage cues to predict deception. The methods associated with this category are, as follows (Conroy, Rubin, & Chen, 2015, 3):

Method	Brief presentation
Linked data	An approach that leverages an existing body of collective human knowledge in order to assess the truth of new statements. The method is based on querying existing knowledge networks, or publicly available structured data (e.g. the Google Relation Extraction Corpus (GREC)). The structured data network of entities is connected through a predicate relationship. This particular method can help develop the applicability of fact-checking methods (Conroy, Rubin, & Chen, 2015, 3)
Social network behaviour	besides content analysis, the use of metadata and telltale behavior of questionable sources can be examined. This method focuses on compiling the inclusion of hyperlinks or associated metadata to

	establish veracity assessments. As an example, centering resonance analysis (CRA), is a model of network-based text analysis, representing the content of large sets of texts by identifying the most important words that link other words in the network (Conroy, Rubin, & Chen, 2015, 3-4)

Table 9. Overview of methods employed for network analysis

Machine Learning (ML) solutions to online disinformation

Complementary to the methods presented above, the detection of false information and fake news can be performed by analyzing multiple types of digital content: images, text data, network data, as well as the credibility degree of the author/source and its reputation (Choraś, Demestichas, Giełczyk, & Herrero, 2021, 1-2), as presented in Figure 1 below. A survey conducted by a team of researchers from the University of Albany on the solutions to address fake news detection through text-analysis and mainstream fake news datasets showed that the state-of-the-art approaches for combating the effects of disinformation through detection can be clustered into five main categories, depending on the methods they use (Parikh & Atrey, 2018, 438):

(1) *Linguistic features based methods* – which extract key linguistic features from fake news and false information, as follows:

a) Ngrams: unigrams[64] and bigrams[65] are extracted from the matrix of words in a certain story. These are most often stored as TFIDF (Term Frequency Inverse Document Frequency) values for information retrieval. TFIDF refers to a numerical statistic that is intended to reflect how important a word is to the document that it is used in (Parikh & Atrey, 2018, 438);

b) Punctuation: using punctuation in an article can help the algorithms for fake news detection to make the difference between false and truthful texts. In this case, the punctuation feature collects eleven types of punctuation, implemented through this detection (Parikh & Atrey, 2018, 438);

c) Psycho-linguistic features: which use the LIWC lexicon (Linguistic Inquiry and Word Count) in order to pick out appropriate proportions of words, allowing the system to determine the tone of the language (e.g. positive/negative emotions), statistics of the text (e.g. word counts), part-of-speech category (e.g. articles, nouns, verbs) (Pérez-Rosas, Kleinberg, Lefevre, & Mihalcea, 2018, 5);

d) Readability: includes the extraction of content features such as the number of characters, complex words, long words, number of syllables, word types, and number of paragraphs (Pérez-Rosas, Kleinberg, Lefevre, & Mihalcea, 2018, 5);

e) Syntax: focuses on extracting a set of features based on CFG (context-free grammar), which are heavily dependent on lexicalized production rules combined with their parent

[64] Model which relies on how often a word occurs, without looking at previous words (Devopedia, 2021).

[65] Model which considers only the previous word to predict the current word (Devopedia, 2021).

and grandparent nodes. Functions in this set are also encoded in TFIDF for information retrieval purposes (Parikh & Atrey, 2018, 439).

(2) *Deception modelling based models* – use the two theoretical techniques described below to convert texts into a set of rhetorical relations connected in a hierarchical tree and identifying the results of rhetorical structure relations:

a) Rhetorical Structure Theory (RST): focuses on capturing the logic of a story in terms of functional relations created amongst different meaningful text units, describing, at the same time, a hierarchical structure for each story (Mann & Thompson, 1988). In accordance with the findings of the authors Victoria Rubin, Nadia Conroy and Yimin Chen (Rubin, Conroy, & Chen, 2015), empirical research confirmed in the last decades that writers tend to emphasise certain parts of their papers so as to express in a more evident manner the main ideas expressed in that article. In this context, the RST uses rhetorical connections to identify, in a systematic manner, the emphasized parts of a text (Parikh & Atrey, 2018, 439);

b) Vector Space Modeling (VSM): used to identify the rhetorical structure relations in the sets resulted after the application of RST. VSM helps at interpreting every news text as vectors in high dimensional space, aspect that requires for the extracted text to be modeled in an appropriate manner for the application of various computational algorithms (Rubin, Conroy, & Chen, 2015). In this context, each dimension of a certain vector space refers to the number of rhetorical relations in a complete set of news reports, representation which provides a simple explanation of a vector space, making it available for further analysis (Rubin & Lukoianova, 2014) (Parikh & Atrey, 2018, 439).

(3) *Clustering based models* – a known method to compare and contrast a large amount of data. For example, the gCLUTO package (Graphical CLUstering TOolkit) runs a large number of data set and sorts a small number of clusters using agglomerative clustering with the k-nearest neighbor approach, clustering similar news reports based on the normalized frequency of relations (Rubin, Conroy, & Chen, 2015);
(4) *Predictive modelling based methods* – develop the ability to make predictions about previously unseen news pieces on the results of a logistic regression process (Rubin, Conroy, & Chen, 2015);

(5) *Content cues based models* – a model based on the ideology of what journalists like to write for users and what are the preferences of users in terms of reading (choice gap), that leverages two different analyses: (I) lexical and semantic levels of analysis (automated methods can be used to extract stylometric features of the text, such as subjective terms, word length etc., further used to differentiate between journalistic formats) and (II) syntactic and pragmatic levels of analysis (the pragmatic function of headlines invokes reference to forthcoming parts in the discourse by making reference to forthcoming parts in the news story. This analysis also covers measuring news sites which have more share activity compared to sites that substantially produce more news content) (Parikh & Atrey, 2018, 440).

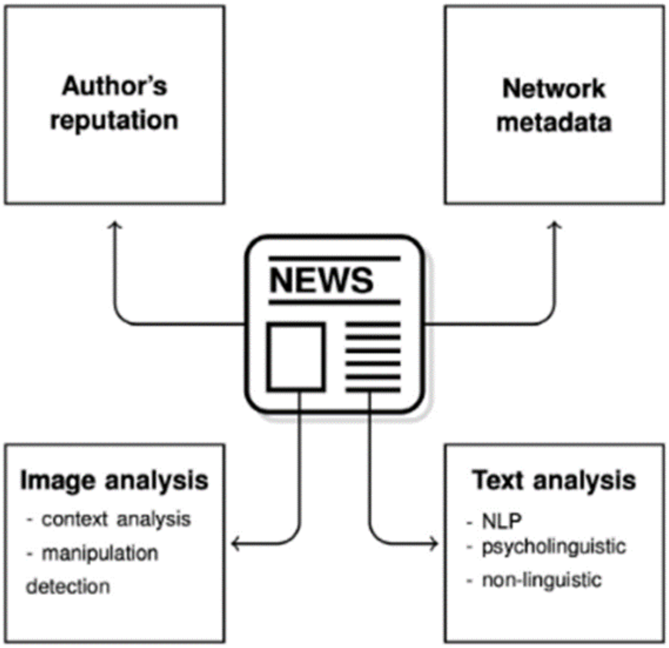


Figure 9. Types of digital content that can be analyzed in order to detect fake news by using automatic instruments (Choraś, Demestichas, Giełczyk, & Herrero, 2021, p. 2)

Considering the solutions briefly described above, we can conclude that there are many Machine Learning initiatives developed in order to combat the negative effects of the online disinformation phenomenon. However, the next section of this chapter will focus on the most common ones, that can be understood by the general public, that does not possess a technological background: (1) serious games – used in order to familiarize the general public with the processes specific to online disinformation and determine it to acknowledge the effects that can be produced by its online behavior and (2) natural language processing tools and (3) social media bot-detecting tools – both used in order to facilitate the process of analyzing the trust level of content disseminated on social media.

However, the chapter does not display a comprehensive analysis of all the technological solutions currently available for combating the effects of online disinformation and propaganda but will only focus on those that were most advertised and that require common, not-specialized skills in order to be used with effective results. Therefore, this chapter will only describe those technological solutions that help to develop digital skills, namely the ones most widely used, examining their potential, but also limitations and societal impact. Lastly, this chapter offers some recommendations for the further improvement of the use of such tools, especially in what concerns compliance with existing human rights standards. However, the chapter will not include a user manual for each example offered but will focus on presenting the main features of the technological solutions and their benefits (the actual cycle of steps that must be followed in order to exploit the features of each application will be developed in the second deliverable of the project).

Inspiring practices, projects, interventions in the field

In addition to the above-mentioned examples, in terms of fake news detection initiatives (with emphasis on rumors), both industry and the scientific community have registered efforts to identify and develop solutions, ranging from research projects (ongoing or already implemented) to fully-fledged applications. The most notable and well-known examples have been collected in the table below (the data has been extracted from Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018, 24-26).

Name of the initiative	Type	Brief description	Site (if available)
PHEME	Project	A 3-year research project funded by the European Commission, implemented during the period of 2014-2017, focusing on the study of natural language processing techniques for dealing with rumour detection and resolution	https://www.pheme.eu/
Emergent	Publication	A data-driven, real-time, web-based rumour tracker, which tracks automatically social media mentions of URLs' associated rumours; however, the identification of rumours and selection of URLs associated with those has not been automated and still requires human input. It was part of a research project led by Craig Silverman, partnering with the Tow Center for Digital Journalism at Columbia University, which focuses on how unverified information and rumour are reported in the media (Silverman, 2015)	www.emergent.info
RumorLens	project	A 1-year research project that was implemented in 2014, funded by Google. Its main objective was to build a tool to aid journalists in	

		finding posts that spread or correct a particular rumour on Twitter, by trying to identify the size of the audiences that those posts have reached. Further details on the rumour detection system developed in this project were published in (Zhao, Resnick, & Mei, 2015)	
TwitterTrails	project	A project in the Social Informatics Lab at Wellesley College. Twitter Trails was developed as an interactive, web-based tool that allows users to conduct an investigation on the origin and propagation characteristics of a rumour and its refutation, if applicable, on Twitter. Visualisations of burst activity, propagation timeline, RT, and co-retweeted networks help its users trace the spread of a story. It collects relevant tweets and automatically answers several important questions regarding a rumour: its originator, burst characteristics, propagators, and main actors according to the audience. In addition, this tool computes and reports the rumour's level of visibility and, as an example of the power of crowdsourcing, the audience's skepticism toward it, which correlates with the rumour's credibility.	http://twittertrails.com/
RumorFlow	application	A framework that designs, adopts and implements multiple visualizations and modelling tools that can be mixed to identify rumour contents and analyze	

		participant activity, either within a rumour, or across different rumours. This approach helps analysts in drawing hypotheses regarding rumour propagation (Dang, Smit, Mod'h, & Minghim, 2016)	
COSMIC	project	A 3-year research project funded by the European Commission, focusing on studying the role of social media to crisis management. In this context, the project studied the adverse use and reliability of social media, including the impact of rumours (Scifo & Baruh, 2013)	
SUPER	project	A 3-year research project funded by the European Commission that studied the use of social sensors for security assessments and proactive emergencies management, dealing also with crowdsourced annotation of rumours (McCreadie, Macdonald, & Ounis, 2015)	
Hoaxey	application	A platform for the collection, detection and analysis of online misinformation and its related fact-checking efforts	http://www.hoaxy.iuni.iu.edu/
Reveal	project	a 3-year project (2013–2016) funded by the European Commission, that aimed at verifying social media content from a journalistic and enterprise perspective, with a focus especially on image verification. The project results were represented by a number of publications on journalistic verification practices concerning	http://revealproject.eu/

		social media (Brandtzaeg, Lüders, Spangenberg, Rath-Wiggins, & Følstad, 2016), social media verification approaches (Andreadou, Papadopoulos, Apostolidis, Krithara, & Kompatsiaris, 2015), and approaches to track down the location of social media users (Middleton & Krivcovs, 2016).	
InVID	project	A Horizon 2020 project, funded by the European Commission (2017-2020), with the target to build a platform providing services to detect, authenticate, and check the reliability and accuracy of newsworthy video files and video content spread via social media.	https://www.invid-project.eu
CrossCheck	project	A collaborative verification project implemented by First Draft and Google News Lab, in collaboration with a number of newsrooms in France, with the objective to fight misinformation (mainly focusing on the French presidential election)	https://crosscheck.firstdraftnews.org /france-en/
Decodex	database	An online database by the French news organization Le Monde, that allows user to check the reliability of news sites	https://www.lemonde.fr /verification/
ClaimBuster	application	A project aiming to perform live fact-checking. The demo application shows check-worthy claims identified by the system for the 2016 U.S. election and it allows the user to input their own text to find factual claims	https://idir.uta.edu/claimbuster/
TwetCred	application	a real-time, web-based system developed to assess the credibility	http://twitdigest.iiitd.edu.in

		of content posted on Twitter. The system does not determine the veracity of stories, but it provides a credibility rating (scored 1 to 7) for each tweet in the Twitter timeline.	/TweetCred/
DeepTrust Alliance	network	a global coalition of stakeholders advancing the fight against digital disinformation and malicious deepfakes	https://www.deeptrustalliance.org/
Unbiased	Project application	a Search Engine which presents insights and analytics on various topics of your choice by combining the power of Crowd-Sourcing with next-generation technologies like Blockchain, Machine Learning and AI.	https://unbiased.cc/search-engine/

Table 10. Overview of fake news detection initiatives

It is clear that there are multiple technologies which play/will play an important role in detecting and combating disinformation. They protect democracies and their citizens from unlawful interference in their internal processes and shed light on the mechanisms used to manipulate public opinions. This positive impact is not without costs. As explained in this deliverable, many of these technologies are still underdevelopment and thus subject to many limitations. In addition to the high error rate, there are also cases where their use can have a negative impact on human rights. In order to avoid this, stronger emphasis needs to be placed on understanding technological limitations, introducing privacy-by-design and privacy-by-default approaches in their developments as well as carrying out a constant review of ways in which their design can be improved in order to mitigate potential risks. In addition to this, it is important to ensure that the regulatory framework manages to keep the pace with technological developments, by introducing the necessary safeguards.

References:

1. Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspective, 31*(2), 211-236.
2. Andreadou, K., Papadopoulos, S., Apostolidis, L., Krithara, A., & Kompatsiaris, Y. (2015). Media REVEALr: A social multimedia monitoring and intelligence system for web multimedia verification. Pacific-Asia Workshop on Intelligence and Security Informatics.
3. Brandtzaeg, P. B., Lüders, M., Spangenberg, J., Rath-Wiggins, L., & Følstad, A. (2016). Emerging journalistic verification practices concerning social media. *Journalism Practice, 10*(3), 323–342.

4. Choraś, M., Demestichas, K., Giełczyk, A., & Herrero, Á. (2021). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing Journal*(101), 1-15.
5. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic Deception Detection: Methods for Finding Fake News. *Proceedings of the Association for Information Science and Technology, 52*(1), 1-4.
6. Feng, V. W., & Hirst, G. (2013). Detecting Deceptive Opinions with Profile Compatibility. Nagoya, Japan: Proceedings of the Sixth International Joint Conference on Natural Language Processing.
7. Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic Stylometry for Deception Detection. Jeju Island, Korea: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).
8. Fowler, G. (2022). *Fake News, Its Impact and How Tech Can Combat Misinformation*. Retrieved January 11, 2023, from https://www.forbes.com/sites/forbesbusinessdevelopmentcouncil/2022/08/22/fake-news-its-impact-and-how-tech-can-combat-misinformation/?sh=308d6bab354f
9. Fraga-Lamas, P., & Fernandez-Carames, T. M. (2020). Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality. *IT Professional*, *22*(2), 53-59.
10. Fuller, C., Biros, D. P., & Wilson, R. L. (2009). Decision Support for Determining Veracity Via Linguistic-Based Cues. *Decision Support Systems*, *46*(3), 695-703.
11. Garimella, K., Gionis, A., Parotsidis, N., & Tatti, N. (2017). Balancing Information Exposure in Social Networks. (pp. 4663-4671). Advances in Neural Information Processing Systems.
12. Kessler, G., Rizzo, S., Kelly, M. (2019, December 16). President Trump has made 15,413 false or misleading claims over 1,055 days. *The Washington Post.* https://www.washingtonpost.com/politics/2019/12/16/president-trump-has-made-false-or-misleading-claims-over-days/
13. Menn, J., Ingram, D. (2017, September 6). Facebook says likely Russian-based operation funded U.S. ads with political message. *Reuters.* https://www.reuters.com/article/us-facebook-propaganda-idUSKCN1BH2VX
14. Middleton, S. E., & Krivcovs, V. (2016). eoparsing and Geosemantics for Social Media: Spatiotemporal Grounding of Content Propagating Rumors to Support Trust and Veracity Analysis during Breaking News. *ACM Transactions on Information Systems*, *34*(3), 1-26.
15. Parikh, S. B., & Atrey, P. K. (2018). Media-Rich Fake News Detection: A Survey. 2018 IEEE Conference on Multimedia Information Processing and Retrieval.
16. Pennycook, G., & Rand, D. G. (2019). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 1-63.
17. Posetti, J., Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. *International Centre for Journalists.* https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf
18. Rubin, V. L., & Lukoianova, T. (2014). Truth and Deception at the Rhetorical Structure Level. *Journal of the Association for Information Science and Technology*, *66*(5), 905-917.
19. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology, 10*(3), 1-42.
20. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explor. Newslett, 19*(1), 22-36.
21. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys, 51*(2), 1-36.

5.2 Serious Games

Alexandra Anghel

Abstract

The present section analyses serious games and the role they can play in assisting people to develop competencies to identify and counter disinformation. Serious games build on the concept of prebunking, previously introduced in section 3.4, and provide an attractive, interactive and hands-on approach to understanding what disinformation is, how it is manifested and what can be done to counter it. However, serious games are not a silver-bullet type of solution and they do exhibit some limitations.

Main research questions addressed

- What are serious games?
- What role do serious games play in countering disinformation?

The concept of serious game is considered to be an oxymoron, given the different domains that the term applies to, from entertainment to education, defense and even healthcare (Djaouti, Alvarez, Jessel, & Rampnoux, 2011, p. 26). Even though the concept of serious game is being considered of recent history, the literature in the field shows that the first use of this term was tracked back in the Renaissance period, where Neo-Platonists used the syntagma of "serio ludere" to refer to the use of light-hearted humor in literature dealing with serious matters (Manning, 2004). A similar idea can also be found in the 1912 Swedish novel entitled "*Den allvarsamma leken*", in English "The Serious Game" (Soderberg, 2001), novel which approaches the delicate topic of adultery. In this context, the "playful" side of cheating is put in opposition with the "serious" consequences of adultery, making the "Serious Game" oxymoron to stress the differences between adultery and the usual definition of games, such as the one provided by the author Johan Huizinga (Huizinga, 1951): "a free activity standing quite consciously outside 'ordinary' life as being 'not serious', but at the same time absorbing the player intensely and utterly" (Huizinga, 1951, 19).

In addition, a similar use of the "serious game" oxymoron was used to describe the professional practice of games and sports. As an example, author Mike Harfield uses this concept in 2008 in his autobiographical book "Not Dark Yet: A Very Funny Book About a Very Serious Game", to refer to his 30-year long career as a professional cricket player (Harfield, 2008). However, the first use of the "serious game" syntagm with a meaning that is close to its current use was traced back to the book "Serious Games" written by the American researcher Clark Abt in 1970 (Abt, 1970), in which he demonstrates how can games be used for training and education.

In order to do this, he designed several computer games such as T.E.M.P.E.R., a game designed for military officers to study the Cold War conflict on a worldwide scale (Raytheon Company, 1965). In addition, in his book, the author also provides examples of "non-digital" serious games, such as math-related games to be used in schools. The definition offered by Abt to the term of serious games, cited by (Djaouti, Alvarez, Jessel, & Rampnoux, 2011) is: "[…] serious games […] have an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement. This does not mean that serious games are not, or should not be, entertaining" (Djaouti, Alvarez, Jessel, & Rampnoux, 2011, 26).

Three years later, a complementary example of a "non-digital" game explicitly labelled as "serious game" is further described by author Donald Jansiewicz in his book "The New Alexandria Simulation: A Serious Game of State and Local Politics" (Jansiewicz, 1973). This book explains the principles of playing a game that was specifically designed to teach the basics of the U.S. political mechanisms. The game was kept in a non-digital format, because the author thought that human interactions are the only ones that can convey the complexity of politics (Jansiewicz, 2011), aspect which allowed it to be used even nowadays in classrooms, in different reissue formats (Djaouti, Alvarez, Jessel, & Rampnoux, 2011, 26-27). Another example of the "serious games" concept used as an oxymoron is the title of an artistic exhibition held in the Barbican Art Gallery from 1996 to 1997, that presented the work of eight artists who sought to make a link between video games and modern art. One of these artists, Regina Corwell, created an interactive art piece in order to ask whether video games can be used as a mean of artistic expression: "If we shift from the fun of games with their overt or covert messages about power, speed, command and control to those same messages delivered for expediency and with urgency by the military and to the efficiency of the office workplace and the various heritage in consumer culture, are art and culture ready to squarely face this complex mosaic?" (Djaouti, Alvarez, Jessel, & Rampnoux, 2011, 27).

This last example limits in a certain way the scope of the concept only to video games, in a similar way to most current definitions offered for the terms of "serious games": "a game in which education (in its various forms) is the primary goal, rather than entertainment" (Michael & Chen, 2006, 17), "serious games have more than just story, art, and software, [they involve] pedagogy: activities that educate or instruct, thereby imparting knowledge or skill. This addition makes games serious. Pedagogy must, however, be subordinate to story—the entertainment component comes first. Once it's worked out, the pedagogy follows" (Zyda, 2005, 26). In fact, all these definitions were influenced by the vision of the author Ben Sawyer expressed in his paper "Serious Games: Improving Public Policy through Game-based Learning and Simulation", published in 2002. The main objective of this paper was to encourage the use of technology and knowledge from the entertainment video game industry to improve game-based simulations in public organizations (Sawyer & Rejeski, 2002).

The above-mentioned paper was the promoter of the "Serious Games Initiative", an association that was founded in 2002 with the aim to promote the use of games for serious purposes (Djaouti, Alvarez, Jessel, & Rampnoux, 2011, 27). Therefore, this is the moment considered to be the "date of birth" of the oxymoron "serious games". In addition, 2002 was also the release date of America's Army, a game considered to be "[...] the first successful and well-executed serious game that gained total public awareness" (Djaouti, Alvarez, Jessel, & Rampnoux, 2011, 27), becoming, as a consequence, the starting point of the serious game current (with the current understanding and use of the concept). Michael Zyda, one of the members of the team that developed America's Army game proposed a definition that is referred to by various research papers: "a mental contest, played with a computer in accordance with specific rules, that uses entertainment, to further government or corporate training, education, health, public policy, and strategic communication objectives" (Zyda, 2005, 26).

Following 2002, most recent definitions of this concept tend to imply the use of digital games, instead of following the broader definition of "serious games" for both digital and non-digital games introduced in the 1970s. (Djaouti, Alvarez, Jessel, & Rampnoux, 2011, 27). One example is the definition used in 2011 by a team of researchers who studied the repurposing of games for educational objectives: "Serious games are very content-rich forms of educational media, often combining high fidelity visual and audio content with diverse pedagogical approaches" (Protopsaltis, et al., 2011, 37).

Taking all the above-mentioned aspects into consideration, we can conclude that the field of serious games had exponentially extended during the last decade, with games being developed in various domains with educational purposes (from the military to the government, education, corporate and healthcare domains). However, even though the literature in the field showed that there were many various definitions of the concept of serious games proposed, none of them succeeded in including all the relevant aspects of the applicability of this term. Therefore, there is no commonly accepted definition of serious game, but all previous, current (and probably) future definitions will focus on the main elements of the concept (as shown in the figure below).

Case studies and lessons learnt

This section will present the main technological driven solutions in terms of serious games and digital initiatives that were developed during the last decade, as mentioned at the beginning of this chapter.

Roozenbeck and van den Linden (Roozenbeek & van der Linden, 2019) designed an experimental game called Bad News!. This game puts people in the position of a person who produces and disseminates fake news. The player's aim in the game is to obtain as many followers and shares as possible through creating and sharing different fake news items. Subjects exposed to the game tended to increase their willingness to engage in critical thinking and to take time to evaluate the accuracy of headlines that researchers exposed them to. After conducting several experiments with a small number of participants (95 in one case and 15000 in another), Roozenbeck and van den Linden (Roozenbeek & van der Linden, 2019) concluded that subjects who had been inoculated against fake news were much more likely to rate fake news as having lower accuracy than real news. This effect was manifested both in cases in which participants were divided into a control group and an experimental group and in the case in which the same people were surveyed before and after playing the Bad News game. A similar result was achieved by Basol, Roozenbeck and van den Linden (Basol, Roozenbeek, & van der Linden, 2021), with an experimental design based on a control group and a treatment group. Lewandowsky and van den

Linden (Lewandowsky & van den Linden, 2021) and Pennycook and Rand (Pennycook & Rand, 2021) summarize the results of the same experiments and show that people who take time to think and evaluate what they see, will be less likely to believe or share disinformation. This leads the researchers to argue that pre-bunking is much more efficient than debunking given that it considerably decreases willingness to believe and share news of a dubious provenance. Van den Linden et al. (van der Linden, Leiserowitz, Rosenthal, & Maibach, 2017) adapted the Bad News game to include information about the COVID-19 pandemic but did not conduct any experiments on the success of this adapted version.

From fake news to chaos! How bad are you? Get as many followers as you can.

Figure 11. Bad News game (Bad News, 2021)

The same authors (Roozenbeek & van der Linden, 2020) also designed Harmony Square, a game that places the player in the shoes of a candidate in elections who can be elected by creating political polarization. This game has also been shown to determine players to rate fake news as less accurate. As described on its website, “the goal of the game is to expose the tactics and manipulation techniques required in order to mislead people, build up a following, or exploit societal tensions for political purposes. Harmony Square works as a psychological “vaccine” against disinformation: playing it builds cognitive resistance against common forms of manipulation that the user may encounter online” (Harmony Square, 2021).

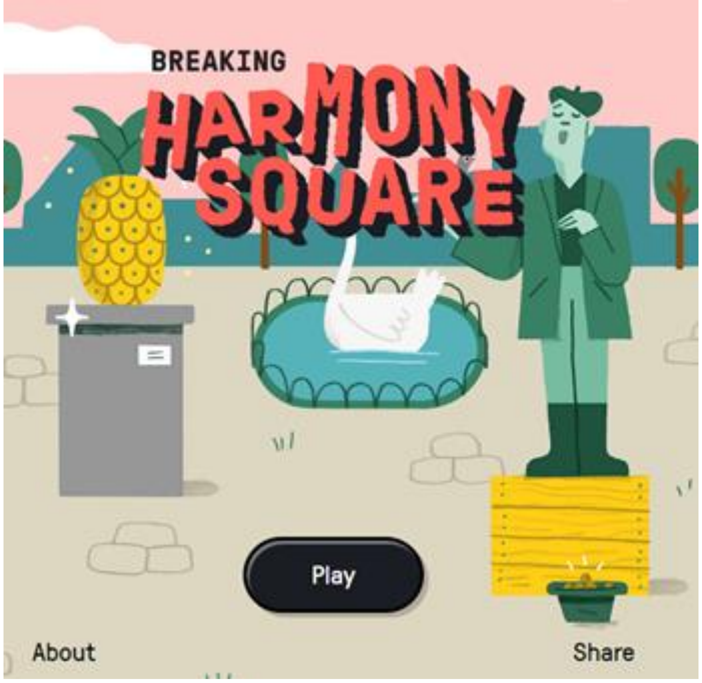
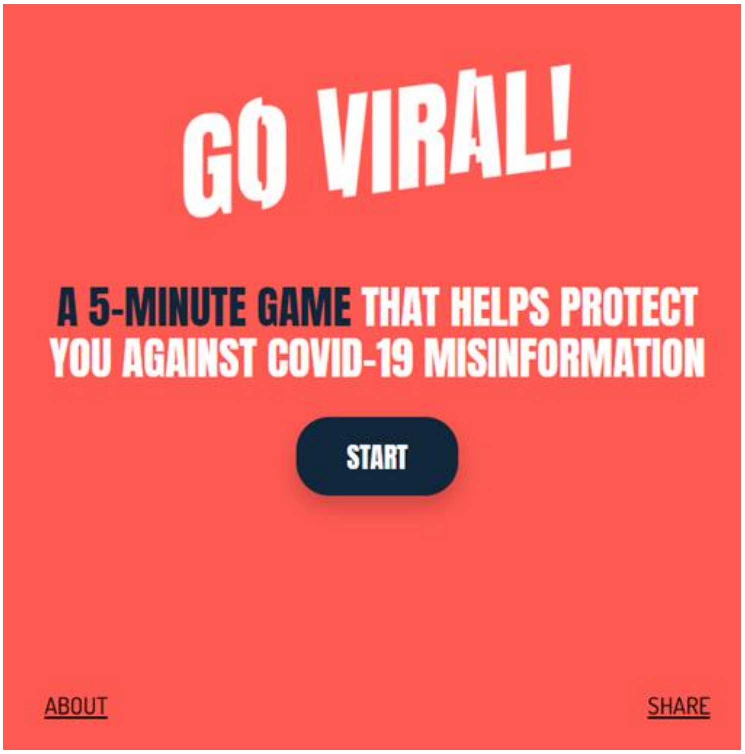


Figure 12. Breaking Harmony Square game (Harmony Square, 2021)

Similar results were obtained by Basol et al. (Basol, Roozenbeek, & van der Linden, 2021), with the game Go Viral, an online game based on misinformation spread during the COVID pandemic. In this particular game, players were asked to imagine that they controlled a social media profile and were asked to obtain as many likes and "credibility points", by sharing posts based on different argumentative fallacies. Basol, et al. (2021) designed several experiments based both on pre-post surveys, as well as on the control-treatment groups design. Further, in the latter experiment, some participants were given an inoculation treatment based on the Go Viral game, others were inoculated with the help of UNESCO infographics about COVID- 19 while others were assigned in the control group. Those who received inoculation treatments were far less likely to believe imaginary social media posts containing demonstrably false information as well as to rate them as highly manipulative.

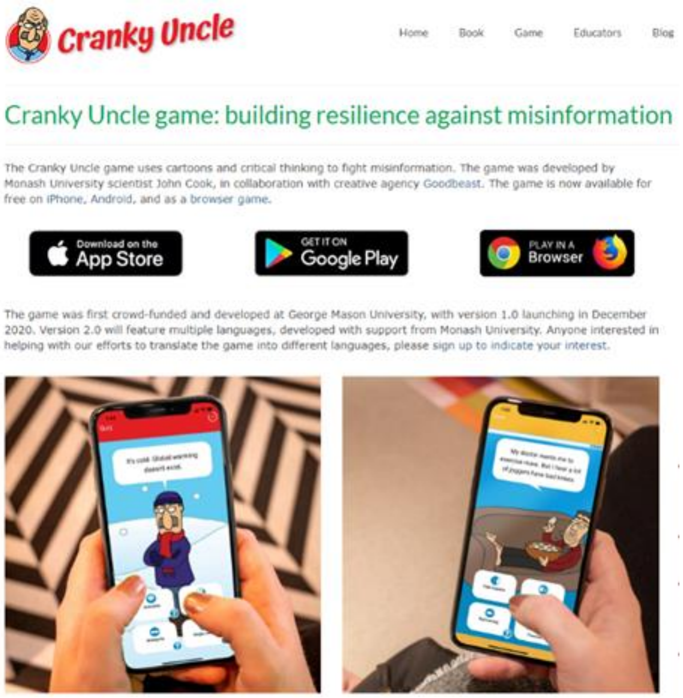


GO VIRAL!

A 5-MINUTE GAME THAT HELPS PROTECT YOU AGAINST COVID-19 MISINFORMATION

Figure 13. Go Viral game (Go Viral, 2021)

Another game identified in the literature was pioneered by Compton et al (2021) and was called Cranky Uncle. This game explains different logical fallacies used by climate change deniers in the form of a cranky old man who issues pronouncements on the non-existence or alternative causes of climate change. Several experiments by researchers showed how playing the game increased the ability to identify and the knowledge of how to use logical fallacies by students in different study programs (Compton, van der Linden, Cook & Basol, 2021).



Cranky Uncle game: building resilience against misinformation

The Cranky Uncle game uses cartoons and critical thinking to fight misinformation. The game was developed by Monash University scientist John Cook, in collaboration with creative agency Goodbeast. The game is now available for free on iPhone, Android, and as a browser game.

The game was first crowd-funded and developed at George Mason University, with version 1.0 launching in December 2020. Version 2.0 will feature multiple languages, developed with support from Monash University. Anyone interested in helping with our efforts to translate the game into different languages, please sign up to indicate your interest.

Figure 14. Cranky Uncle game (Cranky Uncle, 2021)

Another game developed in order to get the users familiarized with the principles of disseminating news in the online environment, as well as with the impact the way in which each piece of news is disseminated can have on the public opinion is the BBC Ireporter. In this game, users “play the role of a social media journalist who is faced with a major breaking story” (Cellan-Jones, 2018). The game is designed to be as realistic as possible, as well as immersive, including elements that involve chatting, having video calls with other journalists and so on. The players need to make decisions with tradeoffs, for example speed and accuracy, whether to publish a story as quickly as possible or to confirm first with a reliable source. The game educates the players more on the side of how good journalism is and what to consider before sharing a story (Bambang, 2020, 4).

Figure 15. BBC Ireporter (BBC, 2019)

In conclusion, serious games could prove a valuable asset in training citizens to detect disinformation attempts and build their resilience against them. Given their playful nature, they could capture the attention of all age groups and make them realise when they are targets of disinformation, thus preventing them from spreading those posts further and contributing to the viralisation of malicious content.

References:

1. Djaouti, D., Alvarez, J., Jessel, J.-P., & Rampnoux, O. (2011). Origins of Serious Games. In *Serious Games and Edutainment Applications* (pp. 25-43). Springer.
2. Hirschberg, J., Manning, C. (2016, May 12). Advances in natural language processing. Science. https://cs224d.stanford.edu/papers/advances.pdf
3. Manning, J. (2004). *The Emblem.* United Kingdom: Reaktion Books.
4. Soderberg, H. (2001). *The Serious Game* (1er edition ed.). United Kingdom: Marion Boyars Publishers Ltd.
5. Huizinga, J. (1951). *Homo Ludens: A Study of the Play-Element in Culture* . Paris: Gallimard.
6. Harfield, M. (2008). *Not Dark Yet: A Very Funny Book About a Very Serious Game.* United Kingdom: Loose Chippings Books.
7. Abt, C. C. (1970). *Serious Games.* New York: Viking Press.
8. Jansiewicz, D. (1973). The New Alexandria simulation; a serious game of state and local politics. San Francisco: Canfield Press.
9. Jansiewicz, D. (2011). The Game of Politics - American Government Simulations. Retrieved January 13, 2023, from https://gameofpolitics.net/faq

10. Michael, D., & Chen, S. (2006). Serious Games: Games that Educate, Train and Inform. Boston, MA: Thomson Course Technology PTR.
11. Zyda, M. (2005). From Visual Simulation to Virtual Reality to Games. Computer, 38(9), 25-32.
12. Sawyer, B., & Rejeski, D. (2002). Serious Games: Improving Public Policy Through Game-Based Learning and Simulation. Washington D.C.: Woodrow Wilson International Center for Scholars.
13. Protopsaltis, A., Auneau, L., Dunwell, I., de Freitas, S., Petridis, P., Arnab, S., . . . Hendrix, M. (2011). Scenario-based Serious Games Repurposing. Pisa, Italy: Proceedings of the 29th ACM International Conference on Design of Communication.
14. Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1-10.
15. Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science advances*, *8*(34), eabo6254.
16. Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1.
17. Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388-402.
18. Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348-384.
19. Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global challenges*, *1*(2), 1600008.
20. van Der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in psychology*, *11*, 566790.
21. Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. V. D. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, *8*(1), 20539517211013868.
22. Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602.
23. Cellan-Jones, R. (2018). *Fake news: Can teenagers spot it?* Retrieved January 15th, 2023, from https://www.bbc.com/news/technology-46206675
24. Bambang, R. J. (2020). *The Fake News Detective: A Game to Learn Busting Fake News as Fact Checkers using Pedagogy for Critical Thinking.* Retrieved January 15th, 2023, from https://smartech.gatech.edu/bitstream/handle/1853/63023/CS6460%20Final%20Paper%20-%20Fake%20News%20Detective%20Game%20-%20rjunior3.pdf?sequence=1&isAllowed=y

5.3 Limits of Technological Tools, Best Practices

Ana Ćuća, Aitana Radu

Abstract

The first part of this section is dedicated to a short introduction on the use of technological tools in the context of preventing and countering disinformation. The section then continues with an analysis of some of these tools, namely the ones most widely used, examining their potential, but also limitations and societal impact. Lastly, this section offers some recommendations for the further improvement of the use of such tools, especially in what concerns compliance with existing human rights standards. This section builds upon the theory presented in deliverable 2.5 Technological factors for disseminating fake news: social media, deepfakes, bots, swarms of bots. It analyses core technological tools used for detecting and countering disinformation and discusses their potential and limitations. It also provides a brief analysis on the human rights impact of such tools.

Main research questions addressed

- Which technological tools are developed for fighting disinformation and what is their potential and limitations?
- What is the human rights impact of such tools?

Tech solutions and their limitations

The majority of tech-driven solutions rely on machine learning. The attractiveness of machine learning in the context of targeting and combating disinformation arises from the fact that machine learning models can recognize novel cases and react to them, based on prior learning. The possibility of continuous improvement of machine learning models, makes them seem like an effective tool to address the always-evolving world of disinformation.

a) Natural Language Processing Tools

Social media content analysis and content moderation have been identified as efficient and effective response instruments to the rising challenge of preventing disinformation. In order to carry out large-scale analysis of social media content, social media companies introduced machine-learning natural language processing tools (Facebook, 2019). Natural language processing (NLP) tools have the ability to parse text "[and] the ability of this paring is usually to predict something about the meaning of the text, such as whether to express a positive or a negative opinion'' (Duarte et al, 2018, p.3). NLP relies on classifiers trained through text labels/annotations determined by humans which guide the tool to decipher whether some word, phrase or text belongs to the targeted category of content. A collection of examples based on which NLP distinguishes different categories of text is called corpus. In the context of detecting disinformation, the NLP tool will use a corpus which has examples of accurate information and disinformation. Disinformation would then be annotated in a way that the tool could learn from this example and employ it automatically in the future. For example, the NLP tool could determine whether some words are missing, but also analyse the word embeddings that represent the context. For NLP tools to analyse the context, they rely on word embeddings generated by machine-learning tools such as Word2Vec (Duarte et al, 2018). NLP tools can replace journalists and media experts in the process of fact-checking. Kozik et. al argue, for example, that NLP tools can imitate a high level of intuitive reasoning, similar to experienced specialists. As computers process large amounts of data, they're able to detect patterns, “without the need to engineer the features before training the neural network” (Kozik et al, 2022). In fact, models rooted in deep-learning can identify authors of fake news based on literary features (Kozik et al, 2022). Example of BENDEMO, Dutch project aiming to prevent and counteract the spread of online disinformation, shows how NLP tools can replace journalist and media experts in the process of fact-checking. Automated network analysis and NLP technology detects emerging disinformation campaigns in the Dutch-speaking region and across Europe, and publishes fact-checks.[24]

Although NLP tools could contribute to a more efficient prevention and countering of disinformation, they are not without limitations and shortcomings. Failing to address these shortcomings, would not only invalidate their efficiency, but quite the contrary they could contribute to spreading more disinformation and/or negatively impacting human rights.

- *Dataset bias*

NLP tools used for combating disinformation are highly dependent on the quality of the training data or in other words, “with limited human direction, an artificial agent is as good as the data it learns from” (Osoba et al., 2017, p.17). This would also mean that, if the data used for training is biased, the automated tool will reproduce these biases, or according to Raso et al. will exacerbate them (Raso et al., 2018). There are several stages in which biases can be introduced in the dataset. In the majority of the cases, bias is introduced during the data collection process, specifically during the annotation process. Individuals building the corpus can integrate their judgement when defining annotations by deciding “what specific type of speech and demographic groups, and so on are prioritized in the training data” (New America, 2020).

An example of dataset bias, that specifically targets one demographic group can be found in a case study put forth by the European Union Agency for Fundamental Rights (FRA). In this particular example, the automatic system indicated a correlation between text being labelled as offensive when written in the African American English dialect, proving that content may be misclassified based on the expressions certain ethnic groups are using (FRA, 2022, p. 69).

- *Language limitations*

Machine-learning NLP tools cannot parse text in all languages. Given the fact that there are hundreds of languages in the world, machine-learning NLP tools will be effective in the case of high-resource languages (HRLs) such as English Spanish, German, and Chinese, while their accuracy will be significantly lower in the case of low-resource languages (LRLs) such as Bengali, Punjabi, Indonesian, although these languages are spoken by millions of people (Hirschberg, Manning, 2016). This would consequently mean that the machine-learning NLP used for detecting disinformation could have "disproportionately harmful outcomes for non-English speakers", especially if the outcome of the machine analysis is to be used as part of a decision-making process (Duarte et al, 2018). According to Xiao Mina, there are a billion people only in Asia who speak thousands of languages and who on the one hand cannot actively participate in conversations in the online world due to language barriers or on the other hand, can experience their posts being flagged as disinformation due to the same language barriers. This would also mean that machine-learning NLP tools would struggle with detecting disinformation in LRLs (Xiao Mina, 2015).

A recently published report by FRA shows how a seemingly neutral corpus covering HRLs, can be biased. During their research, FRA developed several algorithms for offensive speech detection for different languages; English, German and Italian and have found that "for example, in English, the use of terms alluding to 'Muslim', 'gay' or 'Jew' often lead to predictions of generally non-offensive text phrases as being offensive. In the German-language algorithms developed for this report, the terms 'Muslim', 'foreigner' and 'Roma' most often lead to predictions of text as being offensive despite being non-offensive. In the Italian-language algorithms, the terms 'Muslims', 'Africans', 'Jews', 'foreigners', 'Roma' and 'Nigerians' trigger overly strong predictions in relation to offensiveness" (FRA, 2022, p. 11).

- *Accuracy*

The accuracy of these NLP tools is significantly lowered in cases where the context can completely transform the meaning of the claim which the system could mark as disinformation. Since these tools interact with an environment they might not be familiarized with, there is a higher error probability. As Asudeh et al., explain, one may claim that "[she] has never lost a game of chess" which can be truthful information for an experienced chess player, but also for someone who has never played chess (Asudeh et al., 2020). NLP tools have a difficulty distinguishing whether similar claims are truthful or not, since they are entirely context dependent. Furthermore, machine-learning NLP tools experience difficulties in detecting "context, subtlety, sarcasm, and subcultural meaning" (Gillespie, 2020, p. 3). These difficulties were shown in the case of the machine-learning NLP tool used by YouTube which misclassified 150 000 videos as disinformation (Vincent, 2020).

- *Transparency and accountability*

In the words of Frank Pasquale, we live in the Black Box Society in which "hidden algorithms can make (or ruin) reputations, decide the destiny of entrepreneurs, or even devastate an entire economy" (Pasquale, 2016). Given that automated processes such as NLP tools can almost autonomously manage online behaviour as well as enforce rights it is important to apply scrutiny over these and other similar technological solutions. There is often little to no information on how automated tools make correlations or decisions, nor is there any data on their accuracy and reliability. According to Perel et al., "as passive, transparency-driven observations of algorithmic enforcement systems are limited in their capacity to check the practices of non-transparent, constantly evolving algorithms, it is essential to encourage the active engagement of the public in challenging unknown and possibly biased systems…" (Perel, et al., 2017, p.41). As a response to the lack of transparency significant investment has been made in the field of Explainable Artificial Intelligence (XAI) – "a field focused on the understanding and interpretation of the behaviour of AI systems" (Linardatos, 2020, p. 2). Linardatos et. al identified four different methods for explaining algorithms: "methods for explaining complex black-box models, methods for creating white-box models, methods that promote fairness and restrict the existence of discrimination, and, lastly, methods for analysing the sensitivity of model predictions" (Linardatos, 2020, p. 5). Methods for explaining the black-box models aim to interpret already developed complex models, such as deep neural networks (Linardatos, 2020). The white-box models create understandable models which include a decision-tree, making it easier to follow the development process. Methods that promote fairness and restrict the existence of discrimination aim to detect inequities or discrimination which can be promoted by the algorithm. This method is applied with 3 primary goals in mind: controlling discrimination, limiting distortion in individual instances, and preserving utility (Linardatos, 2020, p.19). Lastly, methods for analysing the sensitivity of model predictions aim to ensure that predictions made by the algorithm are trustworthy and reliable (Linardatos, 2020, p. 26).

b) Social media bot-detecting tools

Similarly, to the development of NLP tools, machine learning plays a pivotal role in the detection of social media bots. Detection tools that rely on machine learning can be categorized into three different groups; supervised, unsupervised and semi-supervised. As is the case in machine learning NLP tools, machine learning bot detection tools operate based on the availability and quality of the training data which has been labelled or annotated, specifically providing examples of human-managed and bot-managed social media accounts. Whereas one may argue that in the case of machine-learning NLP tools targeting disinformation, the objective is clear, given that there is at least an EU widely accepted definition, in the context of machine-learning bot detection tools this is more challenging. First and foremost, there is no operational definition of social media bots. Secondly, there is a large grey area in detecting the differences between human-like and bot-like behaviour. Existing bot detection tools rely on datasets that map typical bot behaviour, but the final result is often impacted by the following limitations (Yang et al., 2022):

- *Limited datasets*

The development of supervised social media bot-detecting tools relies on the existence of training datasets. These datasets used for annotation and labelling are often limited, due to them being directly extracted from social media. In recent years, especially in the context of the Cambridge Analytica case, many social platforms have limited access to their APIs as a result of human rights concerns or monetized access making it increasingly difficult to employ social media for training AI models. This means that datasets are compiled by human operators who often manually label and annotate information. Recent research findings seem to indicate a very high error rate in detecting the more sophisticated bots, with only24% of bots being accurately labelled

(Cresci, 2020). The accuracy of training datasets is also impacted by bot evolution. Social media bots were initially easily recognizable since they often lacked personal information and presented few social connections, however with time these bots evolved into perfectly engineered accounts that seem human-operated and displaying a large social network (Cresci, 2020).

- *Language limitations*

According to a 2020 study by Rauchfleisch et al., social media bot-detection tools cannot be transferred from one country to another one. Given the lack of training data available in other languages (all other languages but English), detection tools are likely to give false positives, or negatives since they fail to take into consideration different communication patterns and styles (Rauchfleisch et al, 2020). Even leading instruments such as, Botometer, fail to implement its content and sentiment analysis features in cases of non-English accounts (Yang et al., 2022). This would mean that current machine-learning social media bots detection tools cannot disproportionately target social media users who are non-native English speakers.

- *Misclassification*

Often, human-operated social media accounts can behave similarly to a bot-operated account, especially in the case of politically engaged individuals or activists, namely they don’t disclose a lot of personal data, including their location, and they don’t share any audio-visual content. In other scenarios, individuals try to randomize their handles to protect their data and privacy. Although these are all valid measures individuals might opt to take in order to protect their privacy, due to biases found in the training there is a high likelihood for such accounts to be labelled as bots. Furthermore, taking down such accounts on the premise of them being bots would infringe the freedom of expression of individuals using those social media accounts.

Human rights impact and the way forward

While technology can be an aid in preventing and countering disinformation, it is also clear that many of the tools developed for this purpose can have a negative impact on human rights. For these reasons, the Council of Europe has identified a set of core rights which need to be protected at an individual level, this becoming a key requirement for any technology developed in the field.
1) Technological tools must respect the right to human dignity, the right to life, and the right to physical and mental integrity, defined in Article 2, of the European Convention of Human Rights (ECHR). This would mean that when there is a risk of technological tools violating human dignity, the same procedure must be carried out by a human.
2) The right to liberty and security (Article 5, ECHR), must be respected at all times. This right prescribes an obligation towards developers to establish human rights oversight mechanisms which would evaluate possible risks arising from the implementation of those technologies. Such oversight mechanisms could help in addressing issues arising from dataset bias, language limitations or lack of algorithmic transparency.
3) Special attention must be given to safeguarding the right to non-discrimination (on the basis of the protected grounds set out in Article 14 of the ECHR and Protocol 12 to the ECHR). To prevent

dataset bias, authorities and developers must ensure that deployed technologies do not cause discrimination, promote harmful stereotypes or foster social inequality. Developers must be aware of these risks and continuously examine if in any way bias is fostered through the development and implementation of these technologies.

4) The right to respect for private and family life and protection of personal data (Article 8, ECHR) must be safeguarded. Developers should mitigate any negative impact of technological tools on the right to privacy or family life that might rise either in the development or implementation stage. Protection of this right is particularly relevant in the context of bot detection technologies that tend to show a false positive for profiles where individuals are more protective of their personal information.

5) The right to an effective remedy for violation of rights and freedoms (Article 13 ECHR) must be protected. Authorities and developers should make sure that there are accessible remedies individuals can rely on in case of unlawful data collection or if the implementation of such technologies causes unjust harm to the individual or violates their rights.

6) Similarly, to the above-mentioned right, the right to a fair trial and due process (Article 6, ECHR) should be respected. Individuals must have the opportunity to challenge any decisions made based on evidence acquired through the use of these technologies.

7) Although these technologies are often used to prevent interference in the electoral process through the creation and promotion of disinformation, these tools should in no way inflict the right to freedom of expression and freedom of assembly and association (Article 10 and 11 ECHR). Technologies which target disinformation should respect the principle of transparency, fairness, and responsibility. This obligation is particularly important when it comes to the transparency of algorithms (Leslie et al, 2021).

In conclusion, it is clear that there are multiple technologies which play/will play an important role in detecting and combating disinformation. They protect democracies and their citizens from unlawful interference in their internal processes and shed light on the mechanisms used to manipulate public opinions. This positive impact is not without costs. As explained in this deliverable, many of these technologies are still underdevelopment and thus subject to many limitations. In addition to the high error rate, there are also cases where their use can have a negative impact on human rights. In order to avoid this, stronger emphasis needs to be placed on understanding technological limitations, introducing privacy-by-design and privacy-by-default approaches in their developments as well as carrying out a constant review of ways in which their design can be improved in order to mitigate potential risks. In addition to this, it is important to ensure that the regulatory framework manages to keep the pace with technological developments, by introducing the necessary safeguards.

References:

1. Rauchfleisch, A., Kaiser, J., (2020). The False positive problem of automatic bot detection in social science research. *PLoS ONE*. *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241045*
2. Raso, F., Hilligoss, H., Krishnamurthy, V. et al. (2018, September 25). Artificial Intelligence & Human Rights: Opportunities & Risks. *Berkman Klein Center for Internet & Society, Harvard University.*

3. Posetti, J., Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. *International Centre for Journalists.* https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf
4. Perel, M., Elkin-Koren, N. (2017). Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement. *Florida Law Review*. https://scholarship.law.ufl.edu/cgi/viewcontent.cgi?article=1348&context=flr
5. Osoba O., Wesler IV W. (2017). An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence. *RAND Corporation*. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf
6. New America (n.d). The Limitations of Automated Tools in Content Moderation. *New America*. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-moderation/
7. Duarte, N., Llanso, E., & Loup., A. (2018). Mixed Messages? The Limits of Automated Social Media Content Analysis. *The 2018 Conference on Fairness, Accountability, and Transparency. https://cdt.org/wp-content/uploads/2017/12/FAT-conference-draft-2018.pdf*
8. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*. *https://dx.doi.org/10.3390/e23010018*
9. Kozik, R., Kula, S., Choras, M., Wozniak, M. (2022). Technical solution to counter potential crime: Text analysis to detect fake news and disinformation. *Journal of Computational Science.* https://doi.org/10.1016/j.jocs.2022.101576
10. IBM. (n.d). What is machine learning? *IBM.*https://www.ibm.com/topics/machine-learning
11. European Union Agency for Fundamental Rights. (2022). Bias in Algorithms – Artificial
12. Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*. doi:10.1145/3409116
13. Alaphilippe, A., Gizikis, A. Hanot, C., Bontcheva, K. (2019, March). Automated tackling of disinformation. *European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf*
14. Hirschberg, J., Manning, C. (2016, May 12). Advances in natural language processing. *Science*. https://cs224d.stanford.edu/papers/advances.pdf
15. Pasquale, F. (2016, August 29). The Black Box Society: The Secret Algorithms That Control Money and Information. *Harvard University Press.*
16. Menn, J., Ingram, D. (2017, September 6). Facebook says likely Russian-based operation funded U.S. ads with political message. *Reuters.* https://www.reuters.com/article/us-facebook-propaganda-idUSKCN1BH2VX
17. European Commission. (2018, March 12). Report on the Public Consultation on Fake News and Online Disinformation. *European Commission.* ec.europa.eu/newsroom/dae/document.cfm?doc_id=51810
18. Meta AI. (2019, April 15). Announcing new research awards in NLP and machine translation. *Facebook.* https://ai.facebook.com/blog/announcing-new-research-awards-in-nlp-and-machine-translation/
19. Kessler, G., Rizzo, S., Kelly, M. (2019, December 16). President Trump has made 15,413 false or misleading claims over 1,055 days. *The Washington Post.* https://www.washingtonpost.com/politics/2019/12/16/president-trump-has-made-false-or-misleading-claims-over-days/
20. Asudeh, A. Jagadish, H.V., Wu, Y. Yu, C. (2020, March 11). On detecting cherry-picked trendlines. *Proceedings of the VLDB Endowment.* https://dl.acm.org/doi/10.14778/3380750.3380762
21. Gillespie, T. (2020, July). Content moderation, AI, and the question of scale. *Big Data & Society*. https://www.researchgate.net/publication/343798653_Content_moderation_AI_and_the_question_of_scale
22. Vincent, J. (2020, September 21). YouTube brings back more human moderators after AI systems over-censor. *The Verge.* https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns

23. Leslie, D. Burr, C. et al. (2021, June). ARTIFICIAL INTELLIGENCE, HUMAN RIGHTS, DEMOCRACY, AND THE RULE OF LAW. *Council of Europe.* *https://rm.coe.int/primer-en-new-cover-pages-coe-english-compressed-2754-7186-0228-v-1/1680a2fd4a*
24. Yang, K., Ferrera, E. Menczer, F. (2022, August 20). Botometer 101: social bot practicum for computational social scientists. *Journal of Computational Social Science.* https://link.springer.com/article/10.1007/s42001-022-00177-5

6. PUBLIC POLICY ANALYSIS (WHOLE-OF-SOCIETY APPROACH)

Valentin Stoian-Iordache

Introduction

The chapter analyzes the main policies against disinformation adopted by the European Union and by Romania, Malta and Spain. It identifies two policy models and argues that most actors have chosen an approach based on increasing media literacy competences and on making disinformation more difficult to spread and high-quality information easily accessible. Finally, the section proposes the "whole-of-society" approach as a solution to disinformation, going beyond policies enacted by authorities and empowering actors in society to combat the challenge of disinformation. In the first part, the literature in the field is summarized and a secondary analysis of the relevant policy models is conducted. The literature is presented in a chronological order, focusing on recent academic works in the field. Insights from disciplines such as communication sciences, law and political science are leveraged in order to understand how policy models evolved and were implemented by national and supra-national actors. Secondly, the section presents EU-level policies, with a focus on the Code of Practice against Disinformation and on the several *Communications* issued by the European Commission on the topic. These documents are analyzed in order to highlight the European Commission's approach, focused on a "light touch", in order to avoid actual coercive content removal. The next sections are dedicated to highlighting the national models in three countries: Romania, Malta and Spain. These are presented with the aim of understanding how different countries approached the issue of disinformation in their particular ways, despite being part of the EU and having to follow the same overall legislation. Finally, the theoretical model elaborated by Ivan, Chiru and Arcos (2021), entitled the "whole-of-society approach" is adapted and employed for grounding a solution to the problem of disinformation. While the model was developed for supporting a nation's intelligence efforts, in this case, it is adapted to the national-level effort to combat disinformation.

Digital competences addressed:

1.1 browsing, searching and filtering data, information and digital content;
1.2 evaluating data, information and digital content;
2.3 engaging citizenship through digital technologies.

Main research questions addressed

- What policy models to combat disinformation have been adopted by the EU and by member states?
- What solutions can we propose to address the problem of disinformation?

Saurwein and Spencer-Smith (2020) analyze European and national-level policies against disinformation by using two theoretical frameworks: "assemblage" and "multi-level governance". The ecosystem of disinformation-producing actors is described as a socio-technical assemblage in which producers and sharers of disinformation interact. These have both political and financial motivations. Furthermore, social media users receive and re-share disinformation, while algorithms keep people in the same "content bubble". Policy-makers who engage in the struggle against disinformation are presented as being part of a multi-level governance network, according to the authors.

According to the authors, the EU does not aim to forbid disinformation, but to increase its costs to the point at which it becomes costly to spread it. According to the two authors, the European Commission understands disinformation as a form of market failure, provoked by the fact that "bad information" is "cheap” while "good information" is expensive. Thus, the overall policy direction has been "demonetize bad information" and subsidize high-quality journalism (Saurwein and Spencer-Smith, 2020, 5-6).

Conversely, the authors present the cases of France, Germany and the UK. The first two adopted a "tough approach" including the banning of pieces of disinformation, while the latter undertook a far lighter form of regulation. France introduced a law allowing judges to immediately order the removal of online content if proved to be disinformation. Conversely, the UK established guidelines for potential future regulation. This included imposing a "duty of care" for the platforms. Similarly to France, Germany adopted a tough approach based on the removal of unlawful content (Saurwein and Spencer-Smith, 2020, 9-10).

Mardsen, Meyer and Brown (2020) also discuss EU-level policies against disinformation. They argue in favor of a model called co-regulation which should include actors such as state regulators, civil society and social media providers. This is the only way to ensure, according to the authors, that policies against disinformation are compatible with freedom of speech, as presented in several CJEU decisions, which disallowed AI-enhanced content-based filters. Jason Pielemeier (2020) praises policies undertaken by platforms to decrease the costs of high quality information. This is, according to him, the only way to prevent disinformation in a way compatible with freedom of speech, given that, according to him, sanctioning disinformation under criminal law is almost impossible. Pielemeier (2020, 933-944) also praises the *EU Code of practice on disinformation*, which he presents as a way to determine platforms to self-regulate before any regulation would be imposed on them.

Durach, Bargaoanu and Nastasiu (2020) follow Mardsen, Meyer and Brown's approach (2020) in identifying, four models of regulation: a) self-regulation by platforms; b) co-regulation between supra-national and national authorities, on the one hand, and private actors, on the other,

c) direct regulation and d) audience-centered regulation (or demand-side solutions) that focus on the audience through measures such as increasing media literacy and supporting fact-checking. Audience-based regulation is exemplified by different fact-checking organizations such as EUvsDisinfo, Correctiv (Germany) or Demagog (in Czech Republic) and by different media literacy programs which were promoted through "Media Literacy Week". Similarly to Mardsen, Meyer and Brown (2020), Durach, Bargaoanu and Nastasiu (2020) also defend co-regulation as the best equilibrium between efficiency and legitimacy (for more on media literacy see also 3.3).

Pherson, Ranta and Cannon (2020) employ scenario analysis to present possible futures of the regulation models on disinformation. They rely on two drivers: who initiates the policy (government and private sector) and the type of policy (content-based or user-focused) adopted. The authors elaborate four possible scenarios:

- "Pinocchio warnings" - government-mandated warnings on suspicious content;
- "the Alt-net" - and alternative internet created by the government which one can access only after extensive verification;
- "Rigid gateways" in which Internet providers establish a protocol for verifying content and a standards board;
- "T-cloud", a space which is handled by internet providers where only certified users can post information and which is accessible for a fee.

Hedvig Ördén (2019, 2020) claims that EU policies focused on disinformation have pursued incoherent aims: both a unified narrative, and content pluralism, if this includes only „high quality content". The main reason for this is, according Ördén (2019, 2020), a competition of different epistemic communities: the security/defense establishment (which sees the informational coherence as the key value to be defended) and the the media/journalistic/fact-checker community (which focused primordially on information pluralism as the relevant referent object). Ördén foresees incoherence in the implementation phase, which will give rise to more conflict between these communities.

Case studies and lessons learnt

6.1 Case Study (1) - EU-level policies Valentin Stoian-Iordache

The EU Commission first showed interest in the problem of disinformation after becoming aware of its impact on the political events of 2016 in the US and the UK. The Commission requested the elaboration of a report, which was issued in 2018, and was entitled "A multi-dimensional approach to disinformation". It discusses how disinformation harms a democratic society, presents the measures taken by platforms and proposes five directions of action and policy goals. These could be summarized as increasing transparency in online information, especially by flagging paid or misleading posts, creating a transparency index for sources of information, improving media literacy and helping young adults identify fake news, developing tools for assessing the veracity of information, by empowering journalists and by electronic programs, and creating a diverse media ecosystem without government interference (EU Commission 2018a).

The overarching document which sets out the Commission's policies in the area is the "Tackling online disinformation: a European approach" (EU Commission 2018b) *Communication*. This showcases a number of principles on which the Commission planned to ground its actions and several measures which the Commission planned to adopt in the future. The principles could be summarized as: transparency regarding the origin of the information, diversity of information in the information ecosystem so that citizens can make informed decisions, fostering the credibility of information by showing which sources are trustworthy and fashioning inclusive solutions by increasing media literacy and raising awareness (EU Commission 2018b).

The *Communication* (EU Commission 2018b) and its follow-up report, issued in December 2018 (EU Commission 2018c), presented a series of measures and the way in which they would be implemented afterwards. For example, before adopting the *Code of Practice on Disinformation*, the Commission convened a multi-stakeholder forum which represented the initial meeting of the signatories. The Commission also aimed to create a network of fact-checkers and began workshops with the fact-checking community. It proposed new funding calls to support technological innovation[66] in the area of combating disinformation, such as blockchain and automatic algorithms that separate disinformation (see also Chapter 5 for more information on blockchain solutions). Further, the Commission supported the integrity of the 2019 European Elections by helping electoral authorities exchange good practices, supported media literacy by organizing a week dedicated to it and making it mandatory for states to increase it, and established funding calls to support independent journalists.

In addition to the "Tackling online disinformation: a European approach" *Communication*, the European Commission and the High Representative for Foreign Affairs, jointly adopted the *Action plan against disinformation* in December 2018 and evaluated its implementation in 2019 (EU Commission and HRVP 2018; EU Commission and HRVP 2019). The Action plan was divided into four major pillars: improving the capabilities to detect, analyze and expose disinformation, strengthening response, mobilizing the private sector and raising awareness and improving societal resilience.

The two institutions increased their Strategic Communication capabilities, especially those of the EEAS and implemented a Rapid Alert System in the case of Disinformation, which was especially useful for the case of the Notre Dame Cathedral Fire. The two institutions also held seminars for journalists addressing the topic of disinformation, and improved the communication of EU-level policies on the topic. The Commission also supported the creation of a European Branch of the International Fact-Checking network, launched the Social Observatory for Social Media Analysis and helped national election authorities improve security in the 2019 European Parliament elections.

Another direction of action for the European Commission has been coordinating the self-regulation of social media through the *Code of Practice on Disinformation* (EU Commission

[66] Eunomia (open source solution to identify sources of information), SocialTruth (distributed ecosystem that allows easy access to various verification services), Provenance (intermediary free solution for digital content verification) and WeVerify (content verification challenges through a participatory verification approach)"

2018d). The implementation of the initial version of the code was assessed, revised and issued in an improved form in 2022. The Code relies on five directions of action:

1. improving the scrutiny of ad placements refers to stopping the monetization of fake news by not allowing sites which misrepresent themselves to place ads on the platforms;
2. making political and issue-based advertising more transparent by labeling it as such and showing who funded it;
3. eliminating automated behavior of fake accounts (bots);
4. empowering consumers through making quality content more visible;
5. empowering the research community by making data available and easy to use.

The Commission assessed the implementation of the *Code of Practice* in 2020 and was not satisfied with the progress achieved (EU Commission 2020a). Despite some advances in some areas, such as eliminating misleading advertisements (hundreds of thousands of actions by Google and tens of thousands of actions by Twitter), improved labeling or complete ban of political ads and cracking down on inauthentic behavior, the overall appraisal was rather negative. One of the main criticisms was that platforms evaluated only advertisements that they hosted themselves and not the content of materials shared by users. The Commission was also dissatisfied with the insufficient verification of issue-based advertising and the inadequate use of tools to increase the visibility of high quality news. Finally, the lack of proper cooperation with fact-checkers and insufficient release of data to researchers was negatively appraised (EU Commission 2020a). Further, the Code was seen as lacking clear definitions, which inhibits coordinated actions in ambiguous areas, and insufficient focus on some specific areas such as micro-targeting of political advertising, fairness of access to political advertising (EU Commission 2020a).

Another document, issued specifically for the period of the pandemic, was the *Tackling COVID-19 disinformation - getting the facts right* (EU Commission and HRVP 2020) *Communication.* It focuses on the need for better strategic communication aimed at combating disinformation narratives, enhancing the Rapid Alert system, improving the exchange on best practices on issues such as micro-targeting and on cooperation with the WHO for promoting correct information on the COVID-19 virus. The Communication also grants an important role to the platforms, especially as promoters of correct content on the virus, such as that issued by the WHO and through the promotion of fact-checked opinions on the pandemic. Further, the Commission aims to raise awareness and increase social media literacy, especially through funding Erasmus+ and European Solidarity Corps projects on the issue of disinformation (EU Commission and HRVP 2020).

The updated and improved *Code of Practice against Disinformation* was issued by the European Commission in 2022 (European Commission 2022a) It was signed by 33 social media companies and trade federations (EU Commission 2022b). Building on the previous version of the Code, the reinforced variant included 44 commitments across the five pillars. The novelties include strengthening the demonetisation of disinformation, including better oversight of those buying advertising on the platform, better cooperation with fact-checkers (committing to the reporting of the number of third-party audits of buyers of disinformation), better control of intermediaries buying advertising in the name of other websites, and improved verification of the content of third-party messages, including advertisements, with the aim of removing disinformation.

On the issue of political advertising, no agreement on the definition of political and issue-based advertising could be reached. In the enhanced code, parties committed to cooperate to achieve this definition, and to put in place mechanisms to clearly distinguish political advertising and paid-for content, even if this is then further relayed by individuals by messaging apps. According to the Code, sponsors of political ads must be clearly identifiable and the main directions of the contracts signed with them have to be made public. Further, a repository should be created for these advertisements. Political and issue-based ads places must be archived in a repository which should be made public (EU Commission 2022a).

Platforms also undertook to eliminate a wider array of inauthentic behavior such as malicious deep fakes, hack-and-leak operations, fake accounts and bot-driven amplification, and the use of influencers and to combat AI-generated content through legally acceptable automated verification systems. Finally, signatories agreed to exchange information about malicious practices (EU Commission 2022a).

On the issue of media literacy and critical thinking, platforms agreed to support their development, especially to conduct campaigns highlighting the modus operandi of malicious actors. Moreover, parties to the Code undertook to prioritize quality content and to release information about criteria used to prioritize and de-prioritize information. The verification of the authenticity of digital content through automated tools and a better cooperation with fact-checkers to flag disinformation even through direct access to the platforms was another commitment that signatories undertook. In order to improve the quality of information online, platforms agreed to issue warnings from authoritative sources on pieces labeled as disinformation, but to also include a system to contest abusive flagging. In order to improve cooperation with the scientific community, the Code foresaw the release of anonymized datasets to interested researchers and, in very specific cases, of personal data, but only to researchers who have received security vetting (EU Commission 2022a).

With reference to fact-checking, signatories agreed to financially support independent fact-checking organizations, and to compile and publically issue reports on the way the decisions of fact-checkers were implemented. To better enforce the Code, the Commission will create a Transparency Center and a task force (EU Commission 2022a).

While the issue of information attacks is not mentioned in the Security Union Strategy (EU Commission 2020b), its implementation report (EU Commission 2022c) focuses on the measures taken against Russian propaganda, such as the banning of RT and Sputnik. Finally, the Strategic Compass (European Council 2022) mentions the actions which the EU will take against disinformation: increasing the ability to understand and analyze the threat, imposing significant costs on perpetrators, and helping independent media support itself financially. Then, the Compass envisions the creation of a toolbox to strengthen the Union's strategic communication capacities, especially of CSDP missions abroad, and the increase of the Rapid Alert System.

Eventually, according to the Compass, a data space storing information on all relevant information related incidents will be created (European Council 2022).

In the next section, we will examine the effort made at a national level, in three countries: Spain, Malta and Romania, to counter disinformation.

6.2 Case Study (2) - Spain - Cristina Arribas, Manuel Gertrudix, Ruben Arcos

In addition to the measures coordinated with the European Union, since 2018 Spain has carried out different actions in the fight against disinformation through its institutions and has established permanent coordination mechanisms between the different bodies of the Public Administration, highlighting the Permanent Commission to Combat Disinformation, established in March 2019. The disinformation issue is included in the *National Cybersecurity Strategy* of 2019, which is focused more on malicious actors - state and non-state - than on the study of disinformation as a specific risk. The *National Security Strategy* of November 2021 anticipates the further development of a national strategy to combat disinformation campaigns.

As mentioned in section 4.1., in 2020 Spain also implemented a *Procedure of Action against Disinformation*, which was adopted through a ministerial order. It aims to bring three European documents in Spanish legislation and to implement them at a national level. It assigns the main responsibility for the fight against disinformation to a Permanent Commission established by the Department of Homeland Security, which will coordinate inter-ministerial actions.

The procedure relies on four levels of action, beginning from passive monitoring to detect disinformation attacks, analyzing any suspected incidents, briefing decision-makers on the nature of the attack and deciding on a response and coordinating it through the National Security Council.

Regarding the management within the framework of the European Union, four different levels are implemented:

Level I (national) involves the collaboration with StratCom for identifying and analysing disinformation events, especially those related to Spain;

Level II refers to the exchange of information to support the actions against disinformation campaigns using the RAS system;

Level III concerns exchange of information to support the decision-making process at the political level through the State Secretary of Communication;

Level IV involves decision-making and coordination at the political level of the National Security Council.

The structure of the institutional system addressing disinformation:

1. The National Security Council.
2. The Situation Committee.
3. The Secretary of State for Communication.
4. The Standing Committee against Disinformation.
5. The competent public authorities: Secretary of State for Communication, Presidency of the Government (DSN), National Intelligence Center (CNI), Communication offices of Ministries, and other relevant bodies.

6. The private sector and civil society, the media, digital platforms, academia, the technology sector, non-governmental organizations, and society at large.

6.3 Case Study (3) - Malta - Aitana Radu

As previously discussed in Chapter 4, Article 82 of the Maltese Criminal Code criminalises the spreading of false news, and makes it an offence to "maliciously spread false news which is likely to alarm public opinion or disturb public good order or the public peace or to create a commotion among the public or among certain classes of the public" (Criminal Code (Malta), Art. 82). The offence carries a possible three-month prison sentence[67]. This provision is aligned with Wardle and Derakhshan's definition of disinformation, as there is a requirement of (a) false information, (b) intention (i.e., maliciously), and (c) harm (i.e., to public opinion or good order). Furthermore, the provisions are also partially aligned with the European Commission's definition of disinformation ('false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm'), but lacks the element of economic gain.

Similar provisions can be found in Article 9(1) of the Press Act, which stipulates that "whosoever shall maliciously, by any means mentioned in article 3, spread false news which is likely to alarm the public opinion, or disturb public order or the public peace, or to create a commotion among the public or among certain classes of the public, shall on conviction be liable to imprisonment for a term not exceeding 3 months or to a fine (multa) or to both such imprisonment and a fine: Provided that, if any disturbance ensues in consequence of the offence or if the offence has contributed to the occurrence of any disturbance, the offender shall be liable to imprisonment for a term of not less than one month but not exceeding six months and to a fine (multa)."

Moreover, while at present Malta has no specific internet content blocking/filtering laws, it should be noted that the absolute majority of laws including criminal and civil laws are to a larger extent technologically neutral and, therefore, can be interpreted to include related activities (Camilleri 2021), especially if we consider Article 82 of the Criminal code together with Article 9(1) of the Press Act.

Within the context of the COVID-19 pandemic, the Maltese state carried out various actions to ensure a continuous flow of official information on the evolution of the pandemic:

- employing daily briefings by the Superintendent of Public Health Charmaine Gauci[68], which were transmitted on both television and social media and which included a Q&A session where both journalists present there and people viewing online could ask questions;

[67] It is important however to consider the limitations of this article, namely that it only applies to instances where it is proven that the false news has contributed or is likely to alarm the public. In cases where the outcome is not this one, the individual in question is not considered liable under article 82 of the Maltese Criminal Law.

[68] In the later stages of the pandemic these briefings became less frequent. However, to preserve the Q&A function, one of the leading newspapers in Malta – Times of Malta introduced a dedicated section entitled Ask Charmaine, where people could ask questions related to COVID-19 and vaccines.

- publishing daily information updates on the number of cases/vaccines on the official Sahha page.

In addition to these measures, the government, through the Department of Information, employed a social media campaign, developed by WHO, aimed at flagging disinformation in the context of the COVID-19 pandemic.

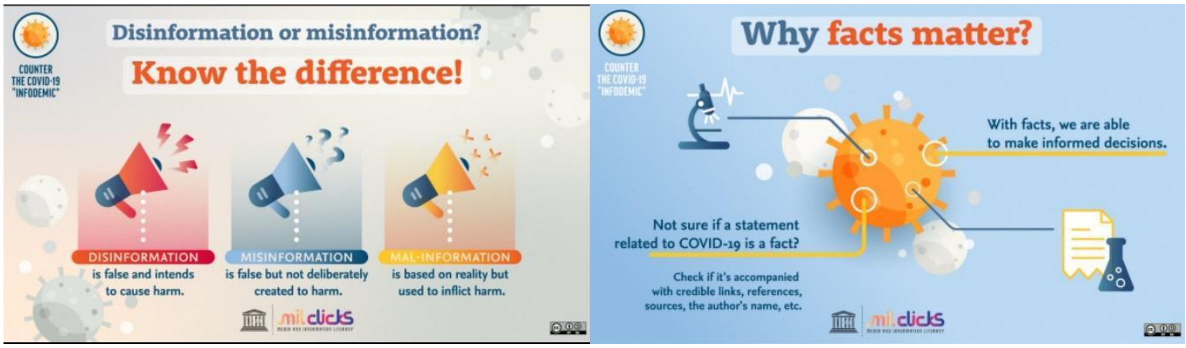

COUNTER THE COVID-19 "INFODEMIC"
Disinformation or misinformation?
Know the difference!
DISINFORMATION is false and intends to cause harm.
MISINFORMATION is false but not deliberately created to harm.
MAL-INFORMATION is based on reality but used to inflict harm.
mil clicks
COUNTER THE COVID-19 "INFODEMIC"
Why facts matter?
With facts, we are able to make informed decisions.
Not sure if a statement related to COVID-19 is a fact?
Check if it's accompanied with credible links, references, sources, the author's name, etc.
mil clicks

Figure 16. WHO social media campaign
Source: https://www.facebook.com/photo/?fbid=2908189345893432&set=disinformation-or-misinformation-know-the-difference

6.4 Case Study (4) - Romania – Valentin Stoian-Iordache

Romania did not adopt legislation allowing the state to prohibit disinformation, with the sole exception of the Decree establishing a state of emergency adopted in March 2020 (Decree 195/2020). Romania declared a state of emergency on the 16th of March 2020, which was in force until the 15th of May the same year (Law 55/2020). Between May 2020 and March 2022, Romania maintained a "state of alert", which included less severe restrictions. The only legislation that expressly allowed authorities to forbid content online and to eliminate websites was in force during the two months of the state of emergency.

Decree 195/2020 permitted authorities to block access to different websites which were considered to spread disinformation. On the basis of this decision, the National Authority for the Administration and Regulation of Communication blocked 15 websites which published disinformation related to COVID-19 (ANCOM 2020). After the end of the state of emergency, there was no legal basis to keep the websites closed and they reopened (Europa libera 2020). No new legal acts were adopted in the wake of the Russian invasion of Ukraine, but Romania directly applied EU regulation 2022/350 of 1 March 2022, and thus banned Russian government-linked websites such as RT and Sputnik.

An EU-funded project entitled "Strategic Planning for consolidating resilience against disinformation and hybrid threats" was launched by the Ministry of Foreign Affairs with the cooperation of the National University of Political Studies and Public Administration. This will create a public policy to strengthen the MFA's ability to combat disinformation in the area of its responsibility (Ministry of Foreign Affairs, 2020).

Romania approaches disinformation as part of the wider concept of hybrid threats. As such, this is mentioned in the *Strategy for National Defence* adopted in 2020. In the *Strategy*, disinformation is described as one of the possible tools employed by enemy actors. According to the document, disinformation can be used to weaken public support for the state's policies. Also, the lack of a robust legal framework for combating disinformation is considered a vulnerability of the Romanian state (Romanian Presidency 2020).

Within the context of the COVID-19 pandemic, the government initiated a public information campaign encouraging people to only get informed from official sources. This was supplemented by several information campaigns on the benefits of vaccination and of wearing masks (Romanian Government 2020). The first campaign featured a video that explained the main characteristics of fake news such as emotionally charged expressions, miracle cures against the coronavirus and the recommendation to share further through social media and messaging apps. The second featured posters showcasing people wearing masks and asking citizens to show that "they care" by wearing masks. Also, the person in the picture was seen addressing the viewers and saying that they are wearing a mask so that "everyone's effort was not in vain" (Timpul 2020). Further, daily briefings by the government presented the number of coronavirus cases and deaths and recommended the actions to be taken by the general population (Ministry of Internal Affairs 2022).

6.5 The whole-of-society approach - a possible solution to the problem of disinformation – Valentin Stoian-Iordache

The policies recently promoted at the European level certainly illustrate the political determination to fight disinformation while integrating and correlating this effort to the strategic communication strategies and plans of the EU. In addition, in different national contexts, proactive and reactive disinformation strategies have been implemented; however, the analysis of different national case studies show significant differences not so much in scope but in pace, reach, systematization, impact, etc.

As it has already been highlighted by the existing debate on the topic, the complexity of disinformation and the rapid and multiple swifts in its manifestations require not just "more initiatives" but anticipatory, innovative and inclusive approaches. Given the specificity of the phenomenon, these cannot be attributed or assigned only to policy makers, media platforms or education providers. Instead, a whole-of-society approach is needed, one that should embrace the principles of multi-stakeholderism (Ivan, Chiru, Arcos 2021), proactive and adaptive behaviors:

- multi-stakeholderism - promoting joint instead of in silo initiatives, collaborative instead of competitive approaches, diverse social, cultural and professional backgrounds instead of traditionally invoked expertise in countering disinformation (e.g. media studies but also first-line practitioners working with vulnerable groups);
- proactive behaviors – anticipating new disinformation trends and investigating the side-effects or unintended outcomes of well-meaning interventions. This must be precluded by a good understanding the vested interests, the actors, and the economic and political mechanisms of disinformation;

- adaptive behaviors – designing measures of intervention based on a cultural and historical awareness of local specificities that may incline people to be more or less skillful in detecting disinformation (e.g. culturally embedded fears and ideas over global power struggles may predispose people to believe in distorted content while also offering them a psychological payoff).

Beyond involving all relevant actors - governmental actors and digital and tech companies, media outlets, civil society organizations (CSOs), citizens, education and research entities, a whole-of–society model of intervention promotes interaction, jointly generated education and knowledge that capacitates society resilience. According to this model, a new framework, including channels to facilitate rapid and impactful communication of relevant information in between state and non-state actors, is created, one in which concurrently:

1. governments encourage independent, professional journalism, avoid censoring content, make online platforms liable for misinformation and fund efforts to enhance news literacy and independent academic research that can inform media intervention and public understanding;
2. media industry focuses on high-quality journalism that builds trust and attracts greater audiences and call out fake news and disinformation without legitimizing them;
3. technology companies strengthen online accountability and invest in technology to find fake news and identify it for users through algorithms and crowdsourcing;
4. educational institutions develop news literacy programs and create opportunities for expert-led independent analysis and business models underpinning this work;
5. citizens take cognitive distance, develop meta-cognitive awareness and dissociative thinking, question institutions and agents of power and protect themselves from false news and disinformation by following a diversity of people and perspectives.

In conclusion, countries have different approaches to the problem of tackling disinformation. As observed in the literature, two main groups of policy models can be discerned: one based on a strong response, including removing content from the internet and another one relying on a softer approach. This involves supporting high quality journalism, empowering fact-checkers and making disinformation more "expensive" to spread. While the European Commission, Romania and Spain opted for the second model, Malta comes closer to the first, even if it did not go as far as Germany or France (fake content is not removed immediately, but only if negative consequences occur.

While policies to combat disinformation have taken many forms, in order to address the phenomenon comprehensively, actors across the whole society should be included. This includes journalists, governments, social media platforms and especially education institutions which need to promote media literacy and critical thinking. This approach, dubbed "whole-of-society," is what the chapter proposes as the only solution wide-ranging enough to meaningfully combat the phenomenon of disinformation.

References:

1. ANCOM 2020, *Decizii ANCOM pentru implementarea prevederilor Decretului nr. 195 din 16 martie 2020 și Decretului nr. 240 din 14 aprilie 2020*, https://www.ancom.ro/decizii-decret-stare-de-urgenta_6253 , accessed 17.08.2022
2. Bennett, A., & Checkel, J. T. (Eds.). (2015). Process tracing. Cambridge University Press.
3. Camilleri, C. (2021). *Regulating disinformation on social media: a European perspective* (Bachelor's thesis, University of Malta), https://www.um.edu.mt/library/oar/handle/123456789/87671, Accessed 8.02.2023
4. Criminal Code, Chapter 9 of the Laws of Malta, available at https://justice.gov.mt/en/pcac/Documents/Criminal%20code.pdf;
5. Decree 195/2020, https://legislatie.just.ro/Public/DetaliiDocumentAfis/223831, accessed 19.08.2022
6. Department of Homeland Security of Spain https://www.dsn.gob.es/es/actualidad/sala-prensa/procedimiento-actuaci%C3%B3n-contra-desinformaci%C3%B3n
7. Durach, F., Bârgăoanu, A., & Nastasiu, C. (2020). Tackling disinformation: EU regulation of the digital space. *Romanian journal of European affairs*, *20*(1).
8. Europa Liberă, *ANCOM: Au fost deblocate toate site-urile închise pentru fake-news*, 15 mai, 2020, accessed 17.08.2022
9. European Commission (2018a), *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union, 2018.
10. European Commission (2018b). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Tackling Online Disinformation: A European Approach COM* (2018) 236 final.
11. European Commission (2018c) *Report from the Commission the European Parliament, the European Council, the European Economic and Social Committee and the Committee of Regions on the implementation of the Communication "Tackling online disinformation: a European Approach*" COM(2018) 794 final
12. European Commission (2018d), *Code of Practice on Disinformation*, https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation, Accessed 25.07.2022
13. European Commission (2020) *Commission Staff Working Document: Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement* 10.9.2020 SWD(2020) 180 final
14. European Commission (2021) *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: European Commission Guidance on Strengthening the Code of Practice on Disinformation* 26.5.2021 COM(2021) 262 final
15. European Commission (2022a) *The Strengthened Code of Practice on Disinformation 2022*, https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation, Accessed 25.07.2022
16. European Commission (2022b) Signatories of the 2022 Strengthened Code of Practice on Disinformation, https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation, Accessed 25.07.2022
17. European Commission and HRVP (2018), *Joint Communication to the European Parliament, the European Council, the European Economic and Social Committee and the Committee of Regions: Action plan against disinformation* 5.12.2018 JOIN(2018) 36 final
18. European Commission and HRVP (2019), *Joint Communication to the European Parliament, the European Council, the European Economic and Social Committee and the Committee of Regions Brussels: Report on the implementation of the Action Plan Against Disinformation* 14.6.2019, JOIN(2019) 12 final

19. European Commission and HRVP (2020), *Joint Communication to the European Parliament, the European Council, the European Economic and Social Committee and the Committee of Regions: Tackling COVID-19 disinformation - Getting the facts right* 10.6.2020, JOIN(2020) 8 final
20. Ivan, C., Chiru, I., & Arcos, R. (2021). A whole of society intelligence approach: critical reassessment of the tools and means used to counter information warfare in the digital age. *Intelligence and National Security*, *36*(4), 495-511.
21. Law 55/2020, https://legislatie.just.ro/Public/DetaliiDocument/225620, accessed 19.08.2022
22. Marsden, C., Meyer, T., & Brown, I. (2020). Platform values and democratic elections: How can the law regulate digital disinformation?. *Computer law & security review*, *36*, 105373.
23. Ministry of Foreign Affairs 2020, "Planificare strategică privind consolidarea rezilienței în fața dezinformării și a amenințărilor de tip hibrid" [Strategic planning on consolidating resilience against disinformation and hybrid threats], https://www.mae.ro/node/55926, accessed 19.08.2022
24. Ministry of Foreign Affairs, European Union, and Cooperation of Spain https://www.exteriores.gob.es/es/PoliticaExterior/Paginas/LaLuchaContraLaDesinformacion.aspx
25. Ministry of Internal Affairs, 2022, Informare COVID-19, Grupul de comunicare strategică [COVID information. Group for strategic communication], 3.03.2022, https://www.mai.gov.ro/informare-covid-19-grupul-de-comunicare-strategica-3-martie-ora-13-00-2/, accessed 19.08.2022
26. National Security Strategy 2021 https://www.dsn.gob.es/es/documento/estrategia-seguridad-nacional-2021
27. Ng, K. C., Tang, J., & Lee, D. (2021). The effect of platform intervention policies on fake news dissemination and survival: an empirical examination. *Journal of Management Information Systems*, *38*(4), 898-930 .
28. Ördén, H. (2019). Deferring substance: EU policy and the information threat. *Intelligence and National Security*, *34*(3), 421-437.
29. Ördén, Hedvig. *Securing Judgement: Rethinking security and online information threats*. Diss. Department of Political Science, Stockholm University, 2020.
30. Order PCM/1030/2020/ Procedure of Action against Disinformation https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-13663
31. Pherson, R. H., Mort Ranta, P., & Cannon, C. (2021). Strategies for combating the scourge of digital disinformation. *International Journal of Intelligence and CounterIntelligence*, *34*(2), 316-341..
32. Pielemeier, J. (2020). Disentangling disinformation: What makes regulating disinformation so difficult?. *Utah L. Rev.*, 917..
33. Press Act, Chapter 248 of the Laws of Malta, available at https://legislation.mt/eli/cap/248;
34. Romanian Government 2020, Get informed only from official sources https://gov.ro/ro/media/video/informeaza-te-doar-din-surse-oficiale&page=1, accessed 17.08.2022
35. Romanian Presidency, *Strategia Națională de Apărare a Țării pentru perioada 2020-2024 [Strategy for National Defense 2020-2024]*, https://legislatie.just.ro/Public/DetaliiDocumentAfis/227499, 2020 accessed 17.08.2022
36. Saurwein, F., & Spencer-Smith, C. (2020). Combating disinformation on social media: Multilevel governance and distributed accountability in Europe. *Digital Journalism*, *8*(6), 820-841.
37. Tangcharoensathien, V., Calleja, N., Nguyen, T., Purnat, T., D'Agostino, M., Garcia-Saiso, S., ... & Briand, S. (2020). Framework for managing the COVID-19 infodemic: methods and results of an online, crowdsourced WHO technical consultation. *Journal of medical Internet research*, *22*(6), e19659.
38. Timpul 2020, "Poartă mască. Arată că îți pasă", 08.07.2020, http://timpul.info/articol/23278/poarta-masca-arata-ca-iti-pasa.html, accessed 13.02.2023