CENTRO ALTI STUDI DIFESA
SCUOLA SUPERIORE UNIVERSITARIA

UNIVERSITÀ DEGLI STUDI DI SALERNO

Dottorato di Ricerca in

Scienze dell'Innovazione per la Difesa e la Sicurezza

XXXVII CICLO

TITOLO DELLA TESI

Countering Harminformation

SETTORE SCIENTIFICO-DISCIPLINARE: INF-01

PRESENTATA DA: FRANCESCO DAVID NOTA

COORDINATRICE DEL DOTTORATO: Prof.ssa Paola Adinolfi

Tutor:
Prof. Nicola Capuano

ANNI ACCADEMICI: 2021/2024

Extended abstract

In the contemporary information landscape, the rapid dissemination of information, particularly on social media platforms, has given rise to the concerning phenomenon of disinformation. The ubiquity and speed of information sharing on social media platforms have amplified the potential impact of disinformation, making it imperative to develop effective strategies for preventing and mitigating its harmfulness. The proliferation of disinformation poses threats to various facets of society, including the erosion of trust in information sources, the manipulation of public opinion, and the potential to incite social and political unrest. The scientific literature has described disinformation as a phenomenon characterised by falseness and intentionality to harm and as the intersection of malinformation (information with the intent to harm but not necessarily false) and misinformation (false information but unintentionally harmful). However, the term has been cause of confusion in the research community because misinformation, disinformation, and malinformation meanings easily overlap as users in social media unintentionally share false information causing harm to the public's opinion. Moreover the term has been used also as a synonym of fake news; however, disinformation does not only involve false news, but also misleading information and manipulated truths. The confusion created in the scientific literature caused by the different use of the same term to indicate different or overlapping meanings, or the association of the same meaning to different terms has inspired the necessity to define the concept of Harminformation. In this thesis the term Harminformation is introduced as a synonym of information harm.

Harminformation is a dynamic phenomena involving several aspects of our lives and societies, a single tool or discipline is not enough to fight it. The cooperation required by experts, institutions and private companies to fight this phenomena has inspired the proposal of an agile framework that offers a conceptual, systematic and process-like approach to counteract it. The framework proposes a structure to formulate, implement and track the most effective short and long term solutions to reduce Harminformation negative impacts. The conceptual framework includes the phases of definition of Harminformation types, identification of solutions, prioritisation of solutions and focus on implementation. In this thesis each step is performed or simulated. The definition phase is performed by extensively reviewing multidisciplinary literature aimed at defining what causes information to be harmful. The definition phase is the description of several existing and newly proposed solutions to each Harminformation type. In this phase metrics characterising in detail what is Harminformation and how to calculate it are proposed. These metrics are based on scores and sub-metrics that characterise one or multiple types of Harminformation, or elaborate characteristics of information typically associated with Harmful consequences according to the scientific literature. The step of Prioritisation of the solutions is simulated, the priority of a solution is calculated starting from the expected velocity of implementation, expected positive outcome and cost. The aim is for institutions and companies to focus first on solutions that are cheaper and are expected to be more effective. The focus phase is described and simulated, together with the plan-do-check-act internal cycle needed for implementation of high priority solutions. The contribution of this work is represented by one of the first attempts in the research community to provide a holistic and clear view of the complex phenomena of information harm starting from the proposal of an agile framework for stakeholders, to the definition of Harminformation, its types and the ways to counteract them. Starting from a multidisciplinary analysis, this work aims to inspire future research and developments toward a standardised process aimed at implementing the most effective solutions to counteract harmful information.

Contents

1. Introduction

1.1 Background

In the contemporary information landscape, the rapid dissemination of information, particularly on social media platforms, has given rise to the concerning phenomenon of disinformation. The ubiquity and speed of information sharing on platforms like X, Facebook, Instagram and Reddit have amplified the potential impact of disinformation, making it imperative to develop effective strategies for preventing and mitigating its harmfulness. The proliferation of disinformation poses threats to various facets of society, including the erosion of trust in information sources, the manipulation of public opinion, and the potential to incite social and political unrest (Pennycook & Rand, 2018; Sunstein, 2017). Disinformation has been proven to affect negatively different aspects of our lives such as politics and elections (Silverman, 2016; Serhan, 2018; Kazeem, 2018; Yaron, 2018; Kušen and Strembeck 2018), racism and fear/hate of ethnic minorities (Müller and Schwarz, 2017; Bursztyn et al., 2018; King's College and Ipsos MORI, 2018), public health (Albarracin et al., 2018; Jolley and Douglas, 2014; Syed-Abdul et al., 2013; Wang et al., 2019), environmental policies (Ward, 2018; Hotten, 2015) economics (Chee, 2020; Bollen et al., 2011) and many others (Valant, 2015; Aisch et al., 2016; Starbird et al., 2014).

In Figure 1 (Hung & Hung 2022) the relationships among the terms Hybrid warfare, Cognitive warfare, Information warfare and Cyber warfare is shown. Hybrid warfare can be defined as a blend of conventional, irregular, and cyber tactics to destabilise opponents; Cognitive warfare targets the human mind, influencing beliefs and decision-making; Information warfare manipulates or disrupts information to gain a strategic advantage. Finally, Cyber warfare uses digital attacks to compromise systems, networks, or data for military or strategic goals.

Being disinformation typically spread through mainstream digital media it can be considered a type of information warfare, which, as shown In Figure 1, is also considered part of Cognitive warfare, and, more in general Hybrid Warfare. Therefore disinformation is a way to influence and control our brain processes. Disinformation deals with media control and intersects with cyber warfare when associated with influence operations.

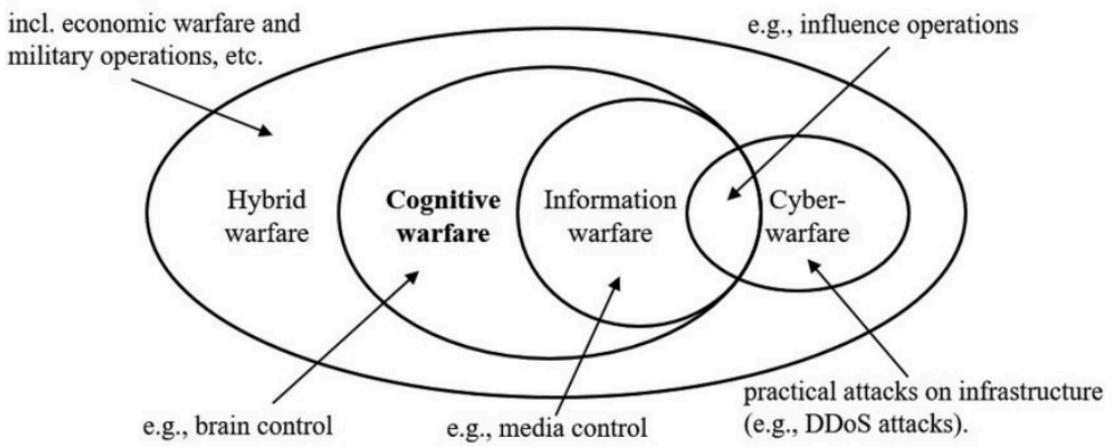

Fig. 1 The Conceptual relationship among cognitive warfare and other types of warfare

As shown in Figure 2 the scientific literature has described disinformation as a phenomenon characterised by falseness and intentionality to harm and as the intersection of malinformation (information with the intent to harm but not necessarily false) and misinformation (false information but unintentionally harmful). However, the term has been cause of confusion in the research community because misinformation, disinformation, and malinformation meanings easily overlap as users in social media unintentionally share false information causing harm to the public's opinion. Moreover the term has been used also as a synonym of fake news; however, disinformation does not only involve false news, but also misleading information and manipulated truths.

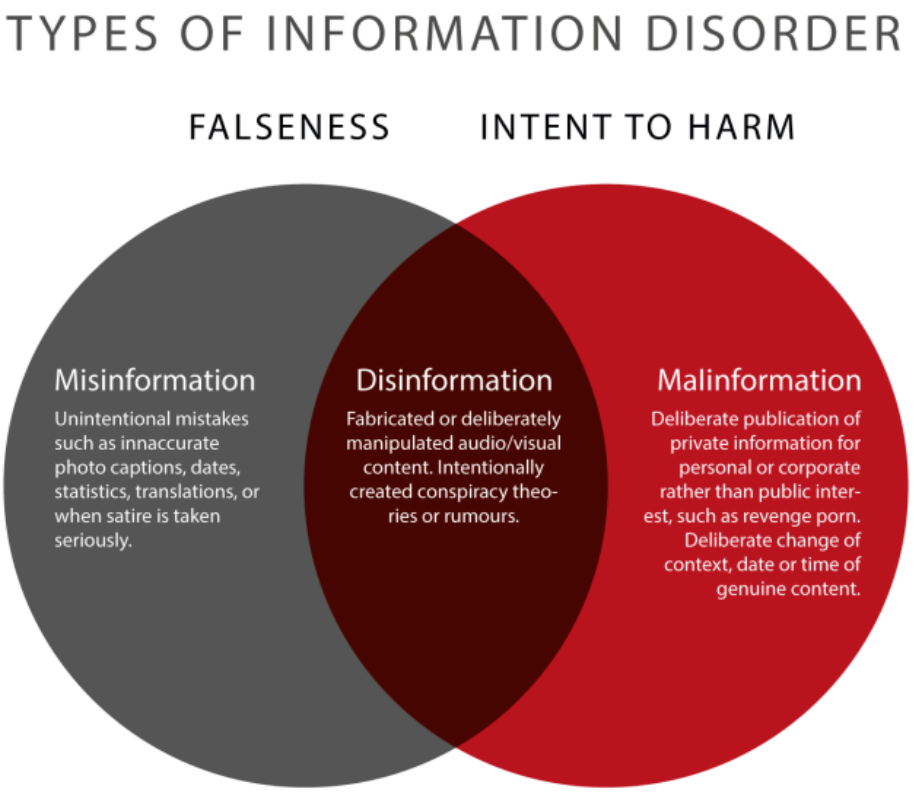

Figure 2 Types of information disorder

The confusion created in the scientific literature caused by the different use of the same term to indicate different or overlapping meanings, or the association of the same meaning to different terms has inspired the necessity to define the concept of Harminformation. Harminformation is a synonym of information that creates harm. The definition of this concept provided in this thesis is abstract, and then, known types of harmful content are discussed to give a precise meaning to the term and avoid interpretations.

Harminformation is a dynamic phenomena involving several aspects of our lives and societies, a single tool or discipline is not enough to counteract it. Artificial intelligence (AI) algorithms used as a form of self-regulation by companies are not enough and usually not adopted unless obliged by clear regulation. On the other hand regulation might be not enough if effective algorithms are not used on the massive amount of data flowing nowadays on social media platforms and news outlets. Regulation must give direction to companies owning social media platforms on the goals that algorithms should

achieve. On the other hand, algorithms should be integrated to identify content that could be harmful and, if necessary, social media platforms should stop harmful information spreading until experts (or accurate AI systems) authorise the diffusion of the content. Although in some cases the idea of harmfulness is clear, such as in pornographic (Flood, 2009) and violent content, in others it is not straightforward thus requiring a clear literature analysis of what is harmful and why.

Most successful Harminformation campaigns target vulnerable individuals, such as ones belonging to social minorities or poor communities; they do so by spreading sensationalism and negativity over a topic during unspecified intervals of time. Identifying these vulnerable communities and applying countermeasures is a critical factor to reduce the harmful effects of these campaigns.

The challenge of reducing harmful content is also one of reducing its spread while, at the same time, balancing content moderation and free speech. In the solutions proposed in this thesis, whenever censorship is a possible negative outcome, strategies to reduce its probability of happening are described.

Existing research highlights the complexity of harmful content dynamics, emphasising the necessity of integrating approaches from multiple disciplines (Starbird et al., 2019; Vosoughi et al., 2018). This thesis aims to contribute to the ongoing discourse on countering it by proposing a conceptual framework that provides a pragmatic process for institutions and organisations to identify, propose and track solutions, while, at the same time, balance the right of free speech and avoid oppressing healthy democratic discussions. The absence of practical, multidisciplinary methods has motivated the solutions suggested in this work. These solutions often require the involvement of various professionals and institutions, including regulators, computer scientists, psychologists, and fact-checkers, who bring diverse expertise.

As highlighted by Broda and Strömbäck (2024) the “solutions” part against harmful content (such as disinformation) is the less studied in the literature; the necessity of analysing and proposing solutions considering the multidisciplinary nature of this problem has never been more important, thus the focus of this work is also to propose real life solutions based on multidisciplinary research analysis and existing tools.

Cognitive scientists, psychologists and computer scientists have studied why people get influenced and why harmful content gets shared; furthermore they categorised types of information that are harmful for society. Solutions to these phenomena must be cost-effective while having a high expected outcome, therefore there is the necessity to provide a way to prioritise them that includes costs, expected positive outcomes and time of implementation.

The contributions of this work are the following: a) the concept of Harminformation is defined and compared to disinformation and other terms; first it is defined in an abstract way and then in a detailed way by listing Harminformation types and metrics to compute its intensity; b) a general framework for stakeholders to counteract harmful content is proposed and described; c) Starting from results from multidisciplinary research the main problems related to harmful content are listed; d) Finally, solutions to these problems are proposed and prioritised with examples of how to implement them and track them.

To my knowledge this is one of the first attempts in the scientific literature to measure and categorise harmful content in a holistic and harmonic way. This attempt aims to reduce confusion created in the

literature by the usage of other terms and shifts the focus from how to define a phenomena to what is harmful and why. This shift should avoid manipulations and interpretations of terms used, for example, in politics to attack opponents. In this work all problems, solutions and metrics are based on this goal. Fighting information harm, fighting Harminformation.

This thesis is organised as follows: section 1 describes the research objectives and significance of this work; In section 2 Disinformation and Harminformation are defined and existing strategies in countering harmful content are reviewed; in section 3 a Framework aimed at representing the organisational processes needed to fight information harm is introduced, together with its structure and its logic; section 4 describes the characteristics of harmful content and the problems identified in the literature, such as how it spreads and why people are unable to detect it; in sections 5,6,7 and 8 several proposals are defined to counteract harmful content and are contextualised on the proposed Agile Framework; section 6 describes in detail the process of prioritisation and simulates it; section 7 discusses the focus step together with the plan-do-check-act-reconsider cycle; finally section 8 includes indications for future research, limitations and conclusions.

1.2 Significance of the study

In an era dominated by information dissemination through social media, the significance of countering harmful content cannot be overstated. The pervasive spread of false, misleading and, more in general, harmful information has profound implications for democratic processes, public trust, and societal stability. As the boundaries between truth and falsehood blur in the online space, there is an urgent need for innovative and comprehensive solutions that address the multifaceted challenges posed by this phenomena. The proposed thesis holds significant implications for both academic research and practical applications. This research addresses the root causes and mechanisms of harmful information spread and damage while suggesting counteracting multidisciplinary solutions.

The societal impact extends to influence public opinion, shape political narratives, and even incite real-world consequences (Allcott & Gentzkow, 2017; Tandoc et al., 2018). Quantitative studies have shown that disinformation causes significant harm in multiple areas.

In the field of public health, the impact of COVID-19 misinformation has been extensively quantified. For instance, Roozenbeek et al. (2020) used surveys and simulations to demonstrate that exposure to COVID-19 misinformation led to a 3% to 12% decline in vaccine acceptance, which directly influenced herd immunity thresholds (Roozenbeek et al., 2020). Similarly, research on vaccination rates has revealed that misinformation often drives vaccine hesitancy. A study by Loomba et al. analysed the correlation between misinformation and declining vaccination rates, showing that in regions where misinformation spread more readily on social media, vaccine uptake was significantly lower, leading to an increased risk of disease outbreaks (Loomba et. al, 2021).

In the sector of politics, studies have focused on how misinformation affects election outcomes and public trust. A 2018 study by Allcott and Gentzkow estimated that fake news circulating during the 2016 U.S. presidential election potentially shifted voting preferences, with fake news having an effect comparable to traditional political campaigning, such as TV ads, influencing as much as 2-3% of the vote in swing states (Allcott and Gentzkow, 2017). Moreover, research by the Pew Research Center has tracked rising distrust in news media and political polarisation, attributing part of this trend to the growing prevalence of disinformation (Jurkowitz et al., 2020). The 2019 Knight Foundation report

found that more than half of Americans exposed to fake news reported feeling greater distrust toward all media, which damaged public trust in institutions (Knight Foundation, 2020).

Economic impacts have also been quantified, particularly in financial markets. A 2020 study by Ferrara et al. analysed how disinformation spread on social media affected stock prices, showing that fake news about publicly traded companies caused statistically significant fluctuations, with stock prices dropping by 5-10% in response to false claims (Ferrara et al., 2016).

As harmful information campaigns become increasingly sophisticated, the need for an agile and conceptual framework becomes imperative (Marwick & Lewis, 2017). As discussed by DiResta et al. harmful content has evolved from a nuisance into high-stakes information war; however, our frameworks for dealing with it have remained the same and focus on counter messaging and counter-narratives, while, at the same time fall into the trap of treating this phenomena as a problem of false stories (DiResta et al., 2020).

In this work several solutions are proposed to counteract harmful information in its complexity and variety. Given the extraordinary amount of news and data being published into nowadays social media it is crucial to perform a preventive and early identification of the news worth to be further analysed, this preselection is beneficial to reduce automated systems processing time, to limit the amount of information fact checkers have to scrutinise and in general to lower down costs. By proposing a framework that focuses on how to plan, implement and track solutions this study also aims to provide a methodology to counteract harmful information.

1.3 Research Objectives

In this section the primary goals of this thesis are outlined. The first aim of this work is in the definition of the concept of Harminformation. This term is introduced as a necessity to avoid the confusion caused by the use of terms such as disinformation, misinformation, malinformation and fake news. For example, disinformation and misinformation are used in different works to represent different or overlapping concepts.

Harminformation includes the concepts of disinformation, misinformation, malinformation, fake news, misleading information, virally overly negative content, overly negative content targeting vulnerable communities, hate speech, violence incitement and prohibited content (such as pornography). This term is adopted throughout all this thesis, although whenever previous research is cited, other terms, such as disinformation and fake news, are also used.

One of the objectives of this thesis is to construct a conceptual framework that offers a systematic and process-like approach to counteract Harminformation. The framework gives the structure to formulate, implement and track short and long term solutions tailored to counteracting harmful information and reduce its negative impacts.

Leveraging insights from crisis communication literature (Coombs, 2014), previous successful interventions (Guess et al., 2019), diffusion models (Vosoughi et al., 2018), short term strategies aim to mitigate the immediate impact of Harminformation. Emphasis will be placed on the agility of response mechanisms to effectively address rapidly evolving scenarios.

Informed by behavioural change theories (Lewandowsky et al., 2012) and experiences in countering disinformation at scale (Starbird et al., 2019), long-term strategies aim to address root causes, enhance digital literacy, and foster a resilient information environment. The goal is to diminish the influence of Harminformation on public trust and societal well-being. Consequently, one expected outcome of the application of this thesis is to contribute to the development of a strategy to counteract Harminformation effectively over time.

The goal of this thesis is to counteract content that is harmful for society. By aligning the proposed framework and solutions with the broader discourse on information disorder (Wardle & Derakhshan, 2017) and incorporating insights from media manipulation research (Marwick & Lewis, 2017), this study aims to offer practical guidance, analyses existing solutions and provides novel proposals to stakeholders including: policymakers, educators, tech companies, fact-checking organisations and national security agencies. These recommendations are intended to strengthen societal defences against harmful content.

Another contribution of this work is the definition of harmful information in all its variants. Specifically, detailed information on almost all types of harmful content are provided.

As research often straddles the realms of theory and practice, a fundamental objective is to bridge this gap effectively. By drawing on communication theories (Sunstein, 2017) and integrating practical experiences from collaborative efforts (Starbird et al., 2019), tangible proposals of solutions to counteract the dynamic nature of harmful content are proposed. These solutions could be, in the long term and with the necessary investments, translated into laws, policies and software to enable institutions (such as the European Commission) or trusted independent organisations to track progress in countering harmful information.

2. Disinformation and Harminformation

2.1 Exploring Disinformation

Disinformation is the intentional spread of false or misleading information with the aim to deceive; alternatively, it is also defined in a report the European Commission by Cock Buning (2018) as content that includes all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit.

This phenomena has become a pervasive challenge with profound implications for society, politics, and public discourse (Wardle & Derakhshan, 2017). Characteristics associated with disinformation highlight the intentional and strategic nature of its creation and dissemination.

Disinformation often exploits the dynamics of online platforms, leveraging algorithms and social network structures to amplify its reach (Zannettou et al., 2019). Characteristics such as emotional appeal, sensationalism, novelty, repetitions and targeting to vulnerable communities contribute to the virality of disinformation, enabling it to spread rapidly within digital platforms and society. Central to this phenomenon is the purposeful distortion of facts, ranging from the creation of entirely fabricated content to the selective presentation of information (Guess et al., 2019). This intentional manipulation of facts contributes to the creation of a false narrative that aligns with the goals of the disinformation campaign. Disinformation campaigns often exhibit strategic timing and targeting to maximise their impact. This involves exploiting current events, crysis, or societal issues to manipulate public perceptions or advance specific agendas (Chen et al., 2020). The intentional targeting of specific demographics or communities is a common characteristic to achieve the desired influence. Attribution of disinformation to specific sources is often challenging due to deceptive tactics, the use of anonymity, and involvement by various state and non-state actors (Hao et al., 2018). Disinformation campaigns may intentionally obfuscate the origin of the information, complicating efforts to trace responsibility.

It has been proven that disinformation in social networks usually targets entire communities; some examples are black communities, veterans, Latin Americans and feminists. The goal is to increase doubts and enforce anxieties to achieve destabilisation of society, political goals, or, in some cases, to gain money and popularity (Marwick et al, 2017).

As discussed by Kahneman (Kahneman, 2017) the human brain has two basic modes of operation: fast and slow. The slow brain takes effort while the fast brain is the efficient part of our brain related to routine. Disinformation attacks the weaknesses of the fast brain. Kahneman has shown that the fast brain can be adapted and changed in the context of disinformation and that developing such an adaptation is key to counteracting it. In recent years information causing surprise, disgust and anger have characterised the spread of disinformation; this can be attributed to the fact that emotions in the fast brain are the first response to new negative information and only after the initial response the slow brain and reason comes into play.

Several studies have shown that emotional content gets more attention, is shared more times and people are less likely to consider whether it is true or not before they share it (Berger & Milkman 2012, Brady et al. 2017). Media literacy, which includes educating people to recognize emotional

content and verify its truthiness in conjunction with social media alerts have been proposed as ideas to stop the spread of it (Rozenbeek et al. 2022).

Disinformation is not primarily a technology-driven phenomenon. The dissemination of false information is mostly driven by socio-psychological factors. Chadwick et al. (2018) report that those who shared tabloid news stories were more likely to share exaggerated or fabricated news. Cognitive psychologists have shown that in fact humans are only 4% better than chance (50%) to distinguish fake from real (Bond and DePaulo, 2006). In Jang and Kim (2018), researchers found that people see members of the opposite party as more vulnerable to false information than members of their party. It is also worth mentioning that people accept more easily information that reflects and reinforces their prior beliefs. This is also known as echo-chambers (Dutton et al., 2017).

In addition to this popular cognition, Pennycook and Rand (2018) suggest that people fall for fake news because they fail to think. Other factors that play a role in deceiving the information consumer are emotions and repetition (Pennycook et al., 2018). Ghanem et al. (2020) showed that each type of false information has different emotional patterns. In their bestseller "Factfulness," Rosling et al. (2018) identify 10 "instincts" such as fear, urgency, and negativity, that lead people in believing false information and developing a distorted view of the world.

As disinformation continues to evolve, researchers must remain vigilant in uncovering and understanding new phenomena and the tactics employed by malicious actors. This comprehension is vital for developing effective countermeasures and strategies to mitigate the societal impact of disinformation.

2.2 Existing strategies to Countering Disinformation

Protective measures against misinformation include educational interventions and evidence-based counter-campaigns. However, implementing these requires addressing several issues: recognizing the severity of the problem, accepting the need to classify information as false/misleading/harmful, and ensuring that interventions respect democratic principles like freedom of expression (Ecker et al. 2024).

Several actions have been conducted to counteract disinformation, including the launch of major counter-disinformation initiatives (European Commission, 2018; Renda, 2018), developing theoretical and computational algorithms, creating educational material ("Bad News Game", 2017), developing fact-checking platforms (InVID Project, 2017; Politifact, 2007; Snopes, 1994), agreeing on a common code of principle for fact-checkers (IFCN, 2017) a self-regulatory Code of Practices for the big social platforms (European Commission, 2018b) and recently the introduction of the Digital Service Act by the European Commission.

In this work the strategies are organised in individual level interventions and system level interventions; this organisation is in line with other works in the literature (Rozenbeek et al. 2022).
In Figure 3 a categorisation of the interventions is represented. Additional detail on each type of intervention is provided in the next sections.

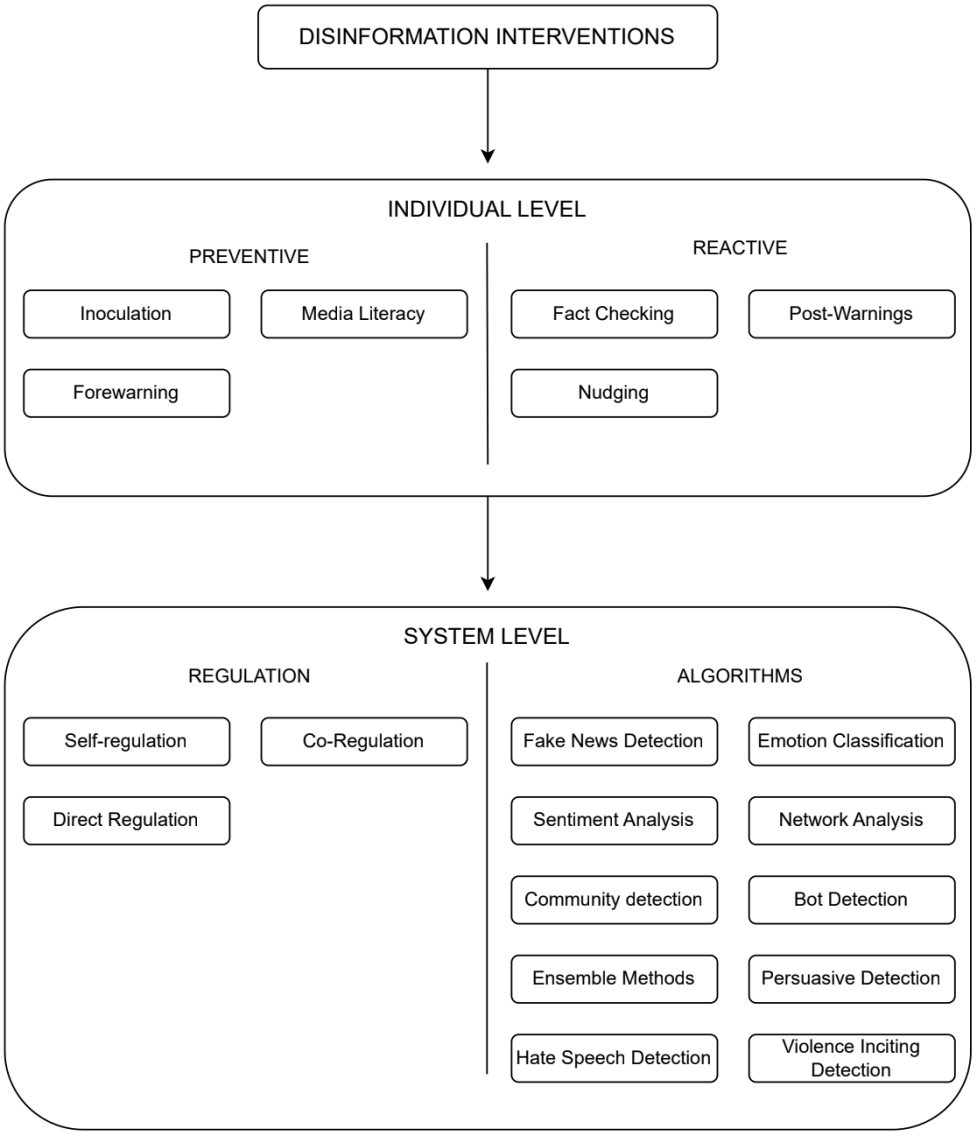

Figure 3 Categorisation of existing disinformation interventions in individual and system level

2.2.1 Individual level interventions

Individual intervention are strategies aimed at changing the individual behaviour and reaction toward disinformation. In this section the following interventions are described: fact checking, media literacy, inoculation, forewarning, post-warnings and nudges. These interventions are either preventive or reactive; preventive solutions aim to stop harmful content from spreading before it affects individuals, while reactive solutions address the issue after the information causing harm has already circulated. Of the interventions described below media literacy, inoculation and nudges are preventive, while fact checking and post-warnings are reactive.

Fact checking is the correction of misinformation (Lewandowsky et al., 2020); it also involves addressing why the misinformation is incorrect and/or providing accurate information (Ecker et al., 2022). Hoes at al. (2024) found that fact-checking, media literacy tips and media coverage of misinformation successfully reduce people's belief in false information, but they also negatively

impact the credibility of factual information. This result is indicative of the inefficacy of fact checking when used as a single countermeasure to disinformation. Furthermore, fact checking usually relies on experts manually checking content; the amount of information created on social media makes it infeasible for manual fact checkers to find and debunk most harmful content. As discussed by Claire Wardle (2017) in a report produced for the council of Europe "While the explosion of fact-checking and debunking initiatives is admirable, there is an urgent need to understand the most effective formats for sparking curiosity and scepticism in audiences about the information they consume and the sources from which that information comes. Simply pushing out more 'factual information' into the ecosystem, without sufficiently understanding the emotional and ritualistic elements of communication, is potentially a waste of time and resources". For example Meta already has a partnership with organisations like PolitiFact or Snopes to identify fake news and remove it from the platform; however, the huge amount of online information makes it difficult to identify fake news at a global scale, in particular when considering disinformation can be spread in many different languages and sometimes includes half-truths which makes it difficult to spot; furthermore, fact checkers do not consider other types of harmful content such as the wide and repetition of negative content against an individual or an organisation. Although manual Fact Checking is an expensive solution it is necessary as expert fact checkers are generally better at integrating contextual information compared to AI algorithms. Individuals may spread false news for several reasons, including to signal political views, or for money, defame opponents and generate chaos; under these circumstances they are less likely to evaluate the veracity of news, and, in some cases they are incentivised in spreading false information (Osmundsen et al., 2021), these reasons are usually related to contextual information rather than specific news, thus the importance of manual fact checking in situations where algorithms are unreliable. Fact checking is a reactive solution as it is possible only when some content has already been shared, this categorises it as a short term solution to counter disinformation.

Media literacy focuses on the ability to evaluate print and online media messages (Potter, 2021). Literacy interventions are provided through formal education or courses (Nygren & Guath, 2021). Several works found media literacy to be effective for improving media literacy skills (Vahedi et al., 2018), media knowledge, and critical perceptions towards media messaging or advertising (Jeong et al., 2012). One of the biggest limitations of media literacy is the necessary cooperation between schools, companies and institutions (such as local and national governments) in order to start media literacy programs. Moreover, these programs tend to have higher costs compared to other solutions and the processes of cooperation, adoption and implementation are slow. These considerations indicate that media literacy is a long term solution to counteract disinformation.

Inoculation is based on the idea of making people more immune to disinformation, exposing individuals to a weakened form of misinformation and/or manipulation strategies to build an ability to resist them. For example, by showing them videos with manipulation techniques commonly used by attackers. Inoculation can be used also after prebuking, that is the strategy of informing people that they will be exposed to disinformation or persuasive attempts in the near future. This strategy is seen as a promising solution in the fight against disinformation as early results show its capability to make individuals more resilient to the influence of harmful content. One limitation of inoculation is the necessity for individuals to participate in the intervention. Furthermore, these interventions create more scepticism toward real news and there is an ongoing debate in the literature on whether this consequence is positive. In order to understand the potential of inoculation more research in real world scenarios is needed; nevertheless, given the early positive results of this technique, its low cost of implementation and its preventive nature it appears to be a good candidate solution to increase

people's resilience against harmful content. Inoculation and Prebunking are currently being used by Google with examples of Prebunking campaigns available online (*Prebunking with Google, 2024*).

Post-warnings involve alerting people that the information they have already been exposed to is false or misleading. It aims to correct misinformation after it has been disseminated. Recently a systematic by Martel et al. (2023) on the efficacy of this method indicates promising results; thus this solution could be considered one of the many necessary solutions needed to counteract harmful content. Post-warnings are usually used after fact checkers or algorithms detect misleading or harmful content.

Nudges are defined as "any aspect of the choice environment that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentive" (Thaler and Sunstein, 2008). Nudging is about asking people to focus on their behaviour while doing an action. In the context of disinformation this translates into some actions the social media platform can have toward a user. Some examples of these actions are: 1) ask a person if the news headline they are reading is accurate; 2) ask if sharing a news without clicking on the link and reading its content is the desired behaviour; 3) ask to focus more time on a piece of news before commenting/sharing it; 4) giving economical incentives to be accurate while writing some content and to flag harmful content. Analysis on nudging effectiveness lead to mixed results, showing accuracy nudges to be ineffective when disinformation is perceived as accurate; however, a field study on X showed that a nudge asking people to share information from higher-quality news sources led to improvements in the quality of the sources people shared (Pennycook et al., 2021). Finally, paying people to be accurate increased discernment in evaluations of news headlines, in particular because people were more likely to identify as correct accurate and true news stories that were incongruent with their beliefs.

2.2.2 System level interventions

System intervention are strategies aimed at changing the systems that are involved, or do not prevent, the spread of disinformation. In this section the following interventions are described: self-regulation, co-regulation, direct regulation and algorithmic interventions. Regulation allows platforms to set and enforce their own rules, enhance collaboration between platforms and governments and use laws or policies imposed by authorities to control harmful information. On the other hand algorithmic interventions focus on creating or using algorithms to reduce the spread of harmful content, flagging or limiting disinformation before it reaches users. Together, these approaches offer different strategies to create a safer information environment.

Self-regulation, as described by Durach et al. (2020), is the idea that digital platforms and online news outlets self-regulate in order to avoid the damage that disinformation does. One of the main weaknesses of the self-regulation approach resides in the conflicts of interest that can emerge, as digital platforms, and especially free-of-charge social networking sites depend on selling as much advertising as possible, as well as on attracting and maintaining the engagement of the audience for as long as possible. Despite these shortcomings, self-regulation is considered to be an important step towards increased accountability of digital platforms. The direct involvement of the digital platforms in countering disinformation is necessary, as actions undertaken by third-party fact checkers are considered insufficient. Conflicts of interest may occur between the platforms' vital necessity to keep users engaged and monetize their engagement, and the public authorities' need to safeguard the integrity and balance of democratic processes within the national public spheres while ensuring the freedom of expression. Other challenges for self-regulation include the enormous amount of content that has to be monitored, the limited efficiency of fact-checking, the occasional failures of human

moderation, and automatic processes of data curation within the digital space. Self-regulation can be useful in all the cases where there are types of disinformation that are platform-specific (such as disinformation that spreads through a feature of a social media that other social media platforms do not offer).

Co-regulation is the cooperation framework among authorities and institutions (such as the European Commission and national-level authorities), the internet platform companies, media organisations, researchers, and other stakeholders. Co-regulation is currently adopted by several institutions. The European Commission, for example, recently approved a regulation called the Digital Service Act (European Commission, 2022). The goal of the European Digital Services Act (DSA) is to add obligations to different kinds of companies to counteract disinformation and harmful content, while increasing transparency and ensuring these obligations provide an environment where every stakeholder can contribute in achieving this goal. For example the transparency Database which lists illegal content notifications received from Very Large Online Platforms (such as Facebook, TikTok, Instagram, YouTube and X) is an example of a Database where researchers can study trends on what content has been regulated on each platform and why. The DSA impact on harmful content and research cannot be analysed at the time of writing of this work as it has been released in early 2024; however, the regulation tackles previously unconsidered problems and makes it easier for institutions to analyse harmful phenomena happening on the different platforms and ask companies to report on how they are taking action to stop them. Companies managing Very Large Platforms (such as TikTok, Meta, Alphabet and X) have interest in counteracting these phenomena to avoid sanctions.

Direct regulation is the practice of taking legal measures and sanctions against companies or individuals spreading harmful content. Within the European Union, Germany and France decided to take stricter measures to counter disinformation, demand more transparency of the digital platforms, apply monetary sanctions, or even block a foreign state-controlled broadcaster. As described by Wagner et al. (2020) the German Network Enforcement Law of 2017, also known as “hate speech law” is addressed to issues like “defamation” or “incitement to crime or violence”. The German law obliges digital platforms with at least two million registered users in Germany to remove illegal content within 24 hours and stipulates fines up to EUR 50 mln, if the content is not deleted. In 2018, a French legislative proposal on the publishing and dissemination of false information during an electoral campaign was enforced. According to the law, an electoral candidate or political party can appeal to a judge to take down a false story or information, within 48 hours. The same law empowers the French broadcasting regulator, the Audio-visual Council, to “block foreign state-controlled broadcasters that publish false information”.

Traditional news media coverage of disinformation has been proposed as a solution to fight disinformation. In the literature there is insufficient evidence on the positive effects of this solution on people’s resilience to harmful content. Although there is evidence of low to none increase in disinformation spreading, unless intentional (Altay et al. 2024). Also media coverage on topics with high uncertainty can exacerbate already difficult situations, for example in a study on COVID-19 news coverage it has been found that right wing news media tend to spread content with low scientific quality, with the consequence of failing to alert the public on health risks related to the virus (Mach et al. 2021).

Algorithmic interventions vary depending on the type of disinformation. The most common types of algorithms and classes of techniques are the following ones:

Fake news detection uses machine and deep learning algorithms to classify news and content as false or true. Typically the classification uses one of the following classes: false, mostly false, mostly true and true. These models analyse linguistic features, metadata, and contextual clues to distinguish between true and false information (Capuano et al. 2023). These algorithms typically work on image (including video) or textual data. Automatic fake news detection remains a huge challenge, primarily because the content is designed in a way that it closely resembles the truth. Fake news tends to have more complicated stories and hardly ever make any references, it is more likely to contain a greater number of words that express negative emotions (Aïmeur et al. 2023). Tanvir et al. (2020) note that it is almost impossible to manually detect the sources and authenticity of fake news effectively and efficiently, due to its fast circulation in such a small amount of time. Therefore, it is crucial to note that the dynamic nature of the various online social platforms, which results in the continued rapid and exponential propagation of such fake content, remains a major challenge that requires further investigation while defining innovative solutions for fake news detection. Researchers concurred that even the best AI for spotting fake news is still ineffective (Hao, 2018). Automatic fake news detection remains a huge challenge, primarily because the content is designed to closely resemble the truth in order to deceive users, and as a result, it is often hard to determine its veracity by AI alone. Most research focuses on studying approaches from a machine learning perspective (Bondielli & Marcelloni 2019; ; Meel & Vishwakarma 2020), data mining perspective (Shu et al. 2017), crowd intelligence perspective (Guo et al. 2020), or knowledge based perspective (Zhou and Zafarani 2020). Although these efforts are necessary they are not enough to counteract a phenomenon that is not mainly technological, but rather psychological and cognitive.

Emotion Classification identifies the emotional tone of content using techniques like lexicon-based approaches, machine and deep learning. Detecting emotions such as fear, anger, disgust and surprise can help flag manipulative or inflammatory messages. Several systems have already been successfully implemented and adopted in the real world. Some examples are GoEmotions-pytorch (Demszky et al. 2020) and HuggingFace emotion recognition (Ravanelli et al. 2021). Some emotion classification algorithms are specialised in classifying the emotions expressed by a face in an image or video. The capability of performing these algorithms on images is crucial to identify also harmful content that is not represented as text online.

Sentiment Analysis evaluates the sentiment (positive, negative, neutral) expressed in content using NLP techniques. Methods include lexicon-based approaches and machine learning models like SVMs, Naive Bayes, and LSTM networks. This helps in identifying potentially harmful or misleading information (Medhat et al. 2014). Sentiment analysis differs from emotion classification as sentiment is only focused on the negativity or positivity of a statement, while emotion classification classifies the intensity of emotions expressed in a sentence.

Network Analysis examines the spread and impact of information across social networks using graph theory and network science techniques. Metrics such as centrality, betweenness, and clustering coefficient are used to identify influential nodes and track misinformation propagation. Some works have shown that in the first stages of content spreading it can be predicted how viral the content will be (Weng et al. 2013); this result and virality predictors are particularly useful to propose some solutions to counteract harmful content. In later chapters some solutions based on these algorithms are cited. Typically these algorithms work on graphs, where each node generally has multiple attributes in the form of textual or numerical data.

Topic Classification uses Natural Language Processing techniques to categorise content into specific topics. Approaches include Latent Dirichlet Allocation (Blei et al. 2003), Non-Negative Matrix Factorization (Lee at al. 2000), and deep learning methods like BERT (Devlin et al. 2018). This helps in identifying unusual or topics. These algorithms are particularly useful when new topics enter the public sphere; in particular they are crucial on topics where there is high uncertainty, such as the Covid-19 pandemic in 2019 when no vaccines and cures were available and insufficient facts were known about how the virus spreaded and its consequences. Later in this work, topic classification is used to calculate a topic-level metric measuring harmful information.

Community Detection analyses social network structures to find clusters of users. Techniques include modularity optimization (Zhang et al. 2009), spectral clustering (Von Luxburg 2007), and community detection methods like the Louvain algorithm (Que et al. 2015). These methods are often helpful in revealing coordinated disinformation campaigns or in identifying communities vulnerable to harmful content.

Bot Detection employs behavioural and network features to flag automated accounts. Techniques include machine and deep learning classifiers and anomaly detection methods (Hayawi et al. 2023). Research efforts have been focused mostly on the platform X given the higher availability of datasets. Therefore, more research is needed in identifying bots on platforms other than X.

Ensemble Methods combine multiple models to improve the accuracy and robustness of disinformation detection. These methods, such as stacking, bagging, and boosting, leverage the strengths of different algorithms to enhance overall performance (Zhang & Ma 2012). Ensemble methods might be of particular interest when models aiming at classifying different characteristics of content, user data and network behaviour are put together to monitor the spread of disinformation. In this work a solution based on this concept is proposed and its pros and cons are analysed.

Persuasive Detection in the context of disinformation is the field of detecting when content, mostly in online discussions, is persuasive. Detecting the persuasiveness of content could be necessary in order to predict the psychological impact of some content on the targeted users (Capuano et al. 2024). This field also includes the study of psychographic profiling, which leverages user data to detect and counteract targeted disinformation efforts by understanding user preferences, behaviours, and personality traits. Techniques include machine learning models that analyse digital footprints, such as social media activity and browsing history (Matz et al. 2017).

Hate speech detection uses AI to detect hate against individuals or groups, in the form of harmful language typically about one or more the following topics: racism, gender discrimination, religion, diseases, radicalization and hurtful language (Malmasi et al. 2017, Zhou et al. 2020, Del Vigna et al. 2017).

Violence inciting detection aims to detect direct violence and passive violence with AI algorithms. Direct violence encompasses explicit threats directed towards individuals or communities, including actions such as killing, rape, vandalism, deportation, desocialization (threats urging individuals or communities to abandon their religion, culture, or traditions), and resocialization (threats of forceful conversion). Passive violence uses derogatory language, abusive remarks, slang targeting individuals or communities and any form of justification for violence (Saha et al. 2023). Few works in research tackle this problem and most of the data analysed is textual. On the other hand most social media companies already have systems and human checkers in place to try to avoid the spread of this type of harmful content on their platforms.

2.3 Defining Harminformation

Harminformation is simply defined as harmful information for individuals and society. The problem of defining what kind of information is harmful has already been previously discussed in the literature. The main characteristics of harmful information are the overload of information (Klingberg 2009), the harm caused by the exposure to redundant and negative content (Pariser 2011, Carr 2010) and high information entropy often spreading higher levels of uncertainty (Stone 2015). Other forms of harmful information are hate speech, fake and misleading news, pornographic and violent content and incitement to violence.

Compared to disinformation and misinformation Harminformation focuses on the harmful consequences of information rather than the intentionality of spreading false and/or misleading content. The need of defining this new term comes from the evident confusion in the usage of terms (such as disinformation, misinformation and fake news) in the literature. Various scientific works and companies use the term disinformation to represent a diverse and overlapping range of concepts.

Although clearly defined, it is often used to characterise one or more types of harmful content: such as disinformation, misinformation, malinformation, infodemic, fake news, hate speech and incitement to violence. Furthermore, the definition of disinformation (the intentional spread of false or misleading information with the aim to deceive) focuses on the intentionality of actors to deceive; if this definition would be adopted as a standard then an individual sharing misleading/false content because he/she believes in it (therefore without intentionality to harm) should be considered as something not worthy of attention and regulation? Would this type of content not be considered harmful?

This confusion of definitions and meanings complicates the work of researchers, companies and institutions in countering harmful content as it is not always clear what the term disinformation refers to. For example disinformation is used in the literature: as a synonym of fake news; as dis- and misinformation; to represent misleading content related to the infodemic phenomena; as forms of hate speech; as a mixed concept including multiple harmful information types.

In other occasions misinformation is used as a term instead of disinformation to represent the same overlapping concepts.One example of overlapping concepts is provided by the reflection described in Humprecht et al. (2020) stating that misinformation, disinformation, and malinformation overlap, as online users unintentionally share false information causing harm to the public's opinion. Humprecht et al. (2020) themselves confuse disinformation with false information, while it has been proven that misleading information is not always false. However, it is crucial to state that not only the intentional spreading of false or misleading content is harmful, and that the non-intentional one might be more harmful for society in some situations.

Recent statistics show that the percentage of unintentional fake news spreaders (people who share fake news without the intention to mislead) over social media is five times higher than intentional spreaders, suggesting the need to counteract non-intentional false news spreading as a priority and avoid referring to fight disinformation as, in principle, it means fighting only intentional spread of misleading/false content.

The definition of Harminformation focuses on the consequence caused by the spread of some content or topic rather than trying to define the type of content itself. Thus, Harminformation can be adopted

to represent: disinformation, misinformation, malinformation, infodemic, hate speech, incitement to violence, viral and overly negative content and overly negative content targeting vulnerable communities. This avoids the confusion created in the literature characterised by the overlapping meanings of these terms as the goal here is to precisely define harmful phenomena. All of these are harmful, but for different reasons, thus the need to define a new term including them all.

The idea of speaking about disinformation and misinformation in terms of harmful content is not new. TrustLab (Cock Buning 2018), a company trusted by the European Commission with the goal of being a leader in moderating content online, is an example of a company often referring to harmful content when speaking about misinformation. Harminformation includes multiple concepts, and is an abstract term, a category composed of all types of harmful content online identified by scientific research.

Oftentimes the scientific literature fall into the trap of treating this phenomena as a problem of false stories (DiResta et al., 2020), harmful content destabilising society is not always based on false news and as the information required to identify something as false might come too late in order to intervene and stop the harmful effects of a Harminformation campaign. Moreover, it is not uncommon for real news to be used maliciously and target social minorities; as detailed by Chadwick et al. (2022), deceptive communication often depends on complex recombinations of true and false information. Bald-faced lies are comparatively rare (McCornack, Morrison, Paik, Wisner, & Zhu, 2014). In some cases establishing the truth of some content is impossible and the harm of real news might be much greater than fake content in some situations.

Balancing the spread of potentially harmful information while keeping healthy democratic discussions is not an easy task. To avoid censorship of legitimate content the solutions to fight Harminformation must be defined together with a metric based on what has been widely accepted as harmful, this is because there is no single way of harming society through information and manipulation on what is harmful and what is not should not be permitted; while a metric based on what scientific literature found to be harmful could reduce censorship of legitimate content. In this thesis a metric is proposed and discussed as a solution to this problem.

The study of solutions to Harminformation focuses on several topics, including: why people are influenced by information, why they share harmful content and how to protect society from harmful content while at the same time ensuring free speech (that includes avoiding censuring negative content that needs to spread to have a functional society). In this work solutions based on established results in the research fields of cognitive science, psychology and computer science are proposed; these solutions come from observations on why Harminformation is effective from an individual and collective point of view.

Harminformation synonyms such as harmful content, information harm and misleading information are used as well throughout this work. In Figure 4 a more detailed representation of the variables influencing Harminformation is shown; these variables are defined in Table 1 and are key to understanding the existing types of Harminformation. Furthermore, in this work a metric based on these variables is proposed.

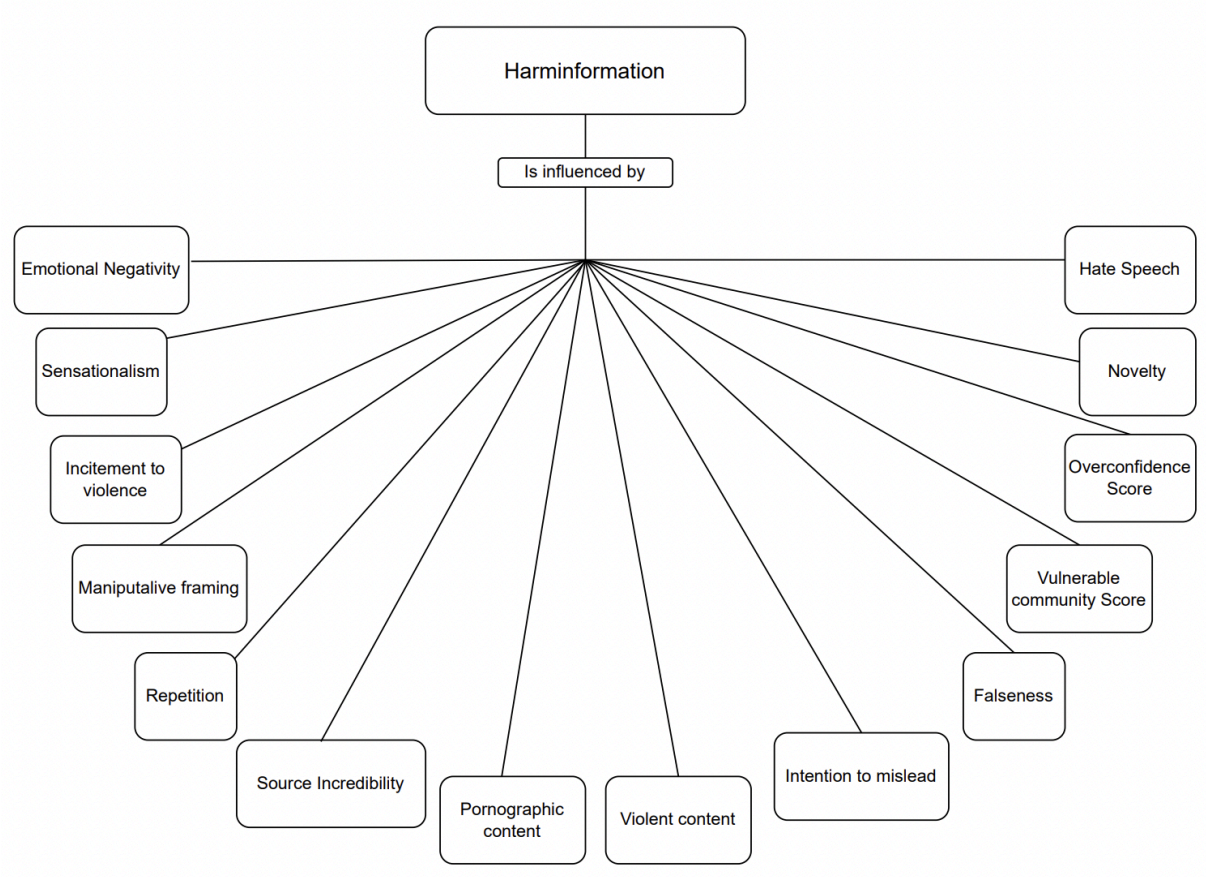

Figure 4 Variables influencing Harminformation

In this thesis this newly defined term is adopted and used to represent a subset of types of harmful content. However, whenever necessary (for example when citing text from other works), other terms are used.

Name	Definition
Emotional Negativity	Measured as the average intensity of emotions like anger, disgust, and fear detected in comments and reactions, using an Emotional Intensity Classifier.
Sensationalism	Assessed based on the presence of exaggerated, shocking, or hyperbolic language in the headline and body text.
Incitement to Violence	Presence and intensity of language encouraging violence.
Manipulative Framing	Use of biased or misleading framing techniques. Such as omitting selectively information.
Repetition	Number of times the content has been shared in a specific timeframe across different sources.
Source Incredibility	Based on a database of known sources and their credibility ratings.
Pornographic Content	Explicit pornographic content.
Violent Content	Violent imagery or descriptions.
Intention to Mislead	Inferred from the context and patterns of Harminformation spread, using behavioural analysis.
Falseness	Verified through fact-checking databases and AI-based fact-checking algorithms.
Vulnerable community score	A score measuring if the content targets vulnerable individuals or communities.
Over confidence score	A score assessing the level of confidence of a text (e.g. of a statement).
Novelty	Measured by comparing the content against a large corpus of similar content to identify if it is anomalous or rarely discussed, using anomaly detection methods.
Hate Speech	Presence and intensity of language that attacks or discriminates against specific groups.

Table 1 Definition of variables influencing Harminformation

3. A framework for countering Harminformation

3.1 Why Agile

The evolving nature of Harminformation campaigns necessitates a dynamic and responsive approach to effectively counteract their impact. Traditional static methods, which rely on pre-existing datasets and predetermined strategies, are insufficient for addressing the fluid and adaptive tactics employed by spreaders of harmful information. An agile methodology, characterised by iterative development, continuous evaluation, and incremental improvement, offers a robust framework for tackling these challenges.

Agile methodologies encourage collaboration among diverse stakeholders, ensuring that solutions are comprehensive and considerate of various impacts. Continuous stakeholder feedback helps refine strategies, making them more resilient to the multifaceted nature of Harminformation. For example institutions such as the European Commission could interact with companies managing very large online platforms such as Meta, Google and TikTok to understand dynamics that are specific to each platform and to favour the transfer of knowledge among companies.

Harminformation campaigns continuously evolve to exploit emerging vulnerabilities and adapt to countermeasures, using sophisticated techniques that render static solutions more obsolete over time. An agile methodology allows for rapid adjustments to these evolving threats, ensuring strategies remain relevant and effective. Solutions to harmful information can rarely be fully controlled using pre-existing databases or controlled environments due to the complexity and unpredictability of real-world scenarios. Agile methodologies facilitate a trial-and-error approach, essential for identifying what solutions work in practice in each context. By deploying interventions in real-world settings and immediately assessing their impact, practitioners can gather critical feedback, learn from failures, and rapidly iterate on their approaches (Highsmith 2009).

In this work an Agile conceptual framework is proposed. The Agile nature of the framework derives from the focus step where the process of planning, implementing, controlling and reacting to a solution should be based on Agile methodologies. The proposed framework suggests the use of the Agile methodology as its principles apply to the constantly changing nature of harmful content in society. However, in this work no detailed guideline is provided to the implementation of the methodology as different solutions and contexts might require variations in the implementation procedures.

3.2 Key concepts and definitions

The issue of counteracting Harminformation is the issue of tackling its multidisciplinary nature. Creating a process that considers, problems, solutions, resources, implementations and measures or reconsiders the progress of them is a necessary step to fight information harm. Therefore, in this work an Agile conceptual Framework is proposed with the aim of guiding progress toward counteracting Harminformation effectively. The framework does not only focus on what solution to implement, but rather on organising the solutions, implementing them and tracking progress effectively.

In Figure 5 the proposed framework is shown. The framework is based on two main cycles, one on the left with a time span of one or more years mostly targeting institutions such as the European

Commission, the one on the right with a timespan of months targeting companies and non-profit organisations counteracting Harminformation.

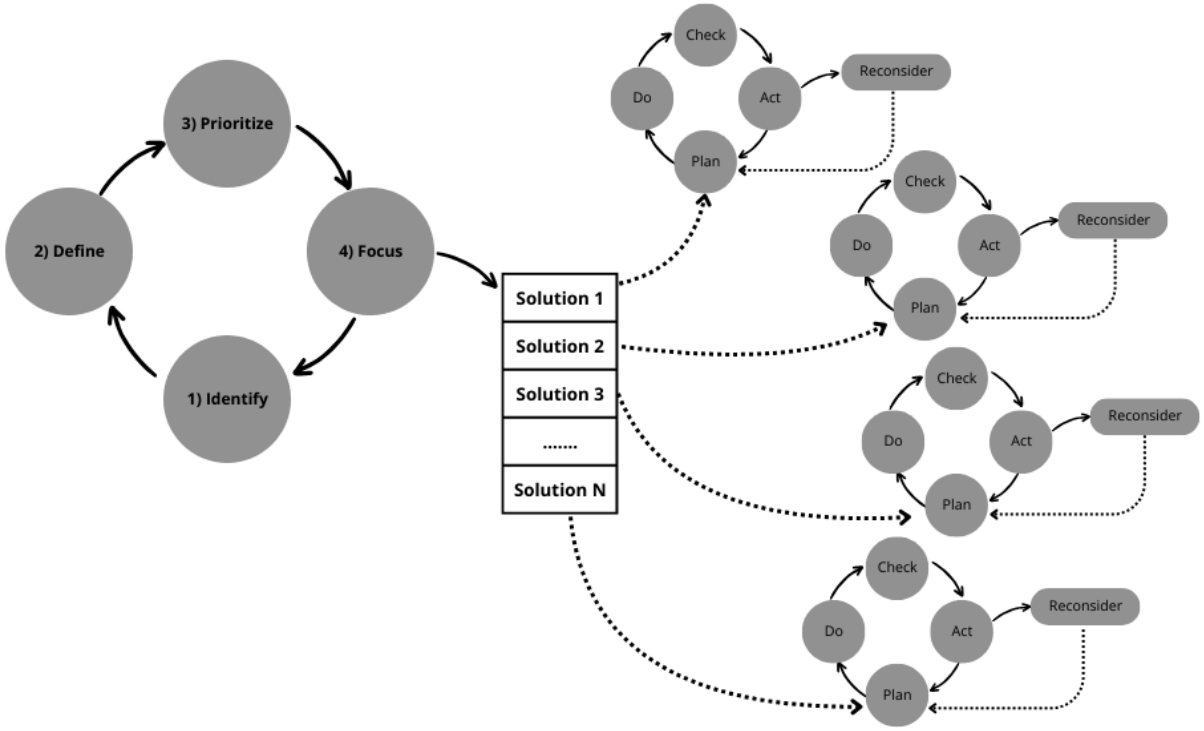

Figure 5 The Agile Framework

The first cycle called Identify-Define-Prioritise-Focus (IDPF) has four steps:

Identify) The first step is to identify problems based on state-of-the-art research on disinformation, misinformation, hate speech and, more in general, harmful content. Research has studied the cognitive phenomena that leads to Harminformation. Depending on the phenomena it is important to define an unresolved problem before giving a solution.

Define) The second step is related to the definition of proposals of solutions. In this phase multiple stakeholders and experts address the problem. Solutions can be proposed by experts of different fields and must refer to the problem identified in step one. The institution receiving the proposals must filter the solutions based on the proponent's credibility and potential conflicts of interests arising from the proposal. Alternatively, if a proponent with conflicts of interest (e.g. a Very Large Company) proposes a solution, then its efficacy will have to be tested and evaluated by independent experts. Depending on the problem, solutions could be proposed by experts in different fields, it is not uncommon to have solutions that require expertise from people of different disciplines. As discussed by Lewandowsky et al. (2012) "to be effective, scientific research into misinformation must be considered within a larger political, technological, and societal context. The post-truth world emerged as a result of societal mega-trends such as a decline in social capital, growing economic inequality, increased polarisation, declining trust in science, and an increasingly fractionated media landscape. We suggest that responses to this malaise must involve technological solutions incorporating psychological principles, an interdisciplinary approach that we describe as technocognition". Although, technology is and will be

one of the main tools to counteract Harminformation the adoption of technology depends on the efforts and investments companies do in counteracting a variety of phenomena; the implementations of solutions can be enforced by increasing regulations in the sector. In this sense, the European DSA is on a good path to counteract some types of harmful information, however, the DSA is vague and does not go into detail of what companies should do to address harmful content, neigher it gives a clear definition of what needs to be counteracted. In this thesis a clear idea on what needs to be counteracted is provided, together with a suggestion on what solutions to use including psychological, technological and legal.

Prioritise) this step consists in prioritising the identified solutions, prioritisation must be based on a voting process. Experts are invited by the institution (e.g. the European Commission) to vote. The invited experts can be members of non-profit organisations (e.g. fact checking organisations), researchers or heads of departments of companies focused on countering Harminformation. Prioritisation has to consider the cost of implementation together with the expected impact. High priority solutions are therefore the ones that will have higher expected positive impact in contrasting Harminformation at the lowest cost and implementation time. The priority is based on the average of scores given by experts. It is important to state that proposed solutions only in this phase must consider the cost of implementation, as different experts could have different ideas of cost and time to implement. However, depending on the sector of implementation some experts might have more weight than others in the prioritisation. For example, if a solution does not include technological considerations in counteracting a problem, experts from social media companies could have a lower weight (e.g. media literacy).

Focus) The fourth step includes the second inner cycle and is related to an extended version of the plan-do-check-act (PDCA) cycle, whose definition and evolution is described by Moen et Al. (2016), and used in several projects as a form of continuous improvement.

In this work the PDCA cycle is adapted to a version composed of five steps (instead of four) and therefore becomes the plan-do-check-act-reconsider (PDCAR):

Plan) In the plan step the institution asks the stakeholders responsible for the problem (e.g. companies managing very large platforms or traditional news media websites) to plan the implementation of the solution;

Do) The stakeholders identified by the institution implement the solution;

Check) The experts must get access to anonymised data provided by the stakeholders that implemented the solution to evaluate its efficacy.

Act) The act step is related to the decision of approval of an implemented solution; the approval depends on reports written by the experts and their opinion on the need for approving the solution, discard it, or adapt it;

Reconsider) This step is introduced in this thesis as an extension of the PDCA cycle. It requires the institution to reconsider an approved solution some years after its adoption (e.g. every two years). This step is necessary as the efficacy of a solution can change over time given the constantly evolving nature of Harmful information.

In this study, the focus phase is simulated to illustrate its practical application. However, further refinement can be achieved by providing concrete examples of how Agile methodology should be structured within the framework of the PDCAR cycle. Given a specific type of harmful information and a proposed mitigation strategy, the PDCAR cycle must be adapted into distinct Agile processes, depending on the organizational and contextual factors influencing the implementation.

For instance, in the case of a software-based intervention developed by a private company, the PDCAR cycle aligns with well-established Agile methodologies commonly employed in the software industry (e.g. Scrum and Kanban). These methodologies can be mapped with the PDCAR cycle in the following way::

Plan: The planning phase involves defining short and medium-term tasks. High-level requirements are outlined for medium-term objectives, while detailed specifications are established for well-defined short-term tasks. Additionally, previously undefined tasks are evaluated, and decisions are made regarding their readiness for design and development. This process is typically led by product owners and product managers, ensuring alignment with strategic goals.

Do: The execution phase encompasses the design, development, and initial testing of the software solution, following the specifications established during the planning phase. Software developers and designers are responsible for implementing the tasks, ensuring the technical feasibility and usability of the solution.

Check: The validation phase assesses the software's behavior in a production environment, with a focus on its efficacy in addressing the identified harmful information issue. Business intelligence analysts and software developers conduct performance evaluations, user feedback assessments, and data-driven analysis to ensure compliance with the intended objectives.

Act: The iterative improvement phase facilitates continuous enhancement of the solution. The effectiveness of the intervention can be refined over time by incorporating variations, optimizing system performance, and mitigating cybersecurity risks. In this context, software developers may propose modifications to improve the solution, engaging in discussions with product managers to integrate these enhancements into future development cycles.

Reconsider: Unlike the previous four steps, reconsideration is not an inherent component of short-term Agile development cycles. Instead, it constitutes a broader strategic analysis to determine whether the conditions and factors contributing to the harmful information problem have evolved, rendering the existing solution obsolete. This phase requires a comprehensive reassessment of both the problem and the implemented solution's effectiveness. For example, a machine learning model trained to detect specific harmful information patterns may become ineffective if adversarial actors adapt their strategies to evade detection. The reconsideration process necessitates in-depth discussions involving diverse stakeholders, including technical leads, product owners, industry leaders, policymakers, and public institutions, to ensure that mitigation strategies remain relevant.

Although the example above outlines the application of Agile methodologies within a private-sector software development context, similar principles can be extended to solutions requiring regulatory interventions from public institutions, citizen engagement, and non-profit organizations. Ideally, the PDCAR cycle should align to the principles of Agile methodologies, regardless of the nature of the implementing entity.

4. Identification phase

The identification step of the framework relies on the idea that new problems arise in time, and, therefore problems related to Harminformation evolve. In this section a literature review, aimed at identifying and describing the different types of harmful content, is performed together with their definition and contextual description.

4.1 The inability to spot Harminformation

Common sense on Harminformation makes us think that bots have an important role in spreading harmful content. However, recent studies have revealed that human users, rather than automated bots, are the primary spreaders of harmful content.

A pivotal study by Vosoughi et al. (2018) found that false information spreads significantly faster, farther, and deeper than true information, primarily due to human engagement rather than bot activity. This is particularly concerning given that users often struggle to discern credible information from falsehoods, especially when presented with emotionally charged content.

Emotional content is a potent catalyst for the spread of Harminformation. According to research by Brady et al. (2017), content that provokes strong emotional reactions such as anger, fear, or empathy tends to be shared more widely and rapidly than neutral content. This effect is amplified in online communities where emotionally charged discourse is prevalent. In particular, minority groups and racially abused communities are particularly susceptible to emotionally manipulative campaigns. These communities can be significantly affected by a mix of real problems, authentic news, false news, and half-truths, making it difficult to discern the veracity of the information they encounter.

The interplay of genuine issues with misleading content creates a fertile ground for Harminformation. As pointed out by Pennycook and Rand (2018), individuals often rely on cognitive heuristics, such as the familiarity of the content or its alignment with pre-existing beliefs, rather than critical analysis to evaluate information. This heuristic process is exacerbated by the emotionally charged nature of the content, leading to a higher likelihood of accepting and sharing false or misleading information.

Furthermore, research by Friggeri et al. (2014) highlights that Harminformation is particularly resilient in echo chambers, where individuals are repeatedly exposed to the same misleading narratives from within their community, reinforcing false beliefs and increasing the difficulty of correction. This effect is particularly pronounced in online environments where algorithmic curation reinforces existing biases and emotional engagement.

Interestingly Landauer (1986) already has shown that adult humans have an estimated memory of around one Gigabyte, which is far less than modern smartphones. Having such a low amount of memory suggests our inability to be good at spotting harmful content as contextual details are often forgotten.

More recent results from Bond and DePaulo (2006) have shown that in fact humans are only 4% better than chance (50%) to distinguish fake from real content.

People often have relevant information stored in memory, but they fail to retrieve and use it under new, incorrect conditions; or described differently, when new claims are false but sufficiently reasonable, people can learn them as facts. For example, when asked “How many animals of each kind did Moses take on the Ark?” Most people responded “two,” even if they knew that Noah, not Mose, built the Ark (Erickson & Mattson, 1981). This phenomenon is called “knowledge neglect” (Marsh & Umanath, 2013). People may also learn the incorrect information and use it in new situations. For example, people who answered the question about Noah’s Ark were more likely to answer the follow-up question “Who built and sailed the Ark?” with “Moses” (Bottoms et al., 2010). This phenomena might explain why it is easy to spread inaccurate and misleading information, in particular when consumed fastly as in social media.

Individuals who reason well with numbers and score high on measures of metacognition (e.g., open-mindedness, reflection vs. intuition) are better at identifying true versus false information (Mirhoseini et al., 2023; Saltor et al., 2023). In a similar study, overconfident people in their ability to distinguish between true and false news were more likely to visit untrustworthy websites and more willing to like or share false content (Lyons et al., 2021).

These works highlight a common result: humans are not good at spotting harmful content; in particular, people are not good at spotting it when it is misleading, false, emotionally charged or when it reinforces our beliefs and comes from communities we belong to.

4.2 Fast and slow thinking

Fast and slow thinking, concepts popularised by Daniel Kahneman in *Thinking, Fast and Slow* (2011), refer to the dual processes of cognition. Fast thinking (System 1) is automatic, quick, and often based on heuristics and emotions, whereas slow thinking (System 2) is deliberate, analytical, and rational. In the context of social media, users predominantly rely on fast thinking due to the rapid and continuous influx of information. This environment fosters fast thinking as users are bombarded with brief, high-frequency updates that demand immediate responses. As a result, decisions about the veracity of information are made hastily, often based on superficial cues such as headlines, images, or the number of likes and shares.

Nicholas Carr, in *The Shallows: What the Internet Is Doing to Our Brains* (2010), argues that the internet encourages a mode of thinking that is less focused and more superficial. Carr explains that the nature of online browsing, characterised by skimming and jumping from one piece of information to another, weakens the ability to engage in deep, reflective thought.

The velocity at which information is disseminated on social media exacerbates the reliance on fast thinking. Users are often overwhelmed by the sheer volume of content, a phenomenon known as information overload. This problem can lead to the rapid spread of Harminformation, as emotionally charged or sensational content is more likely to be shared without critical evaluation.

Users, influenced by cognitive biases such as the availability heuristic (where people judge the likelihood of events based on how easily examples come to mind) and confirmation bias (the tendency to search for, interpret, and remember information in a way that confirms one's preconceptions), are prone to accepting and sharing misleading information. The rapid consumption and dissemination of content leave little time for the verification of facts, allowing false information to propagate quickly.

The interplay between fast and slow thinking on social media platforms significantly impacts how users process and share information, facilitating the spread of harmful content. The preference for fast thinking, driven by the rapid and voluminous nature of social media content, facilitates the spread of harmful content. Understanding this dynamic is crucial for developing strategies to mitigate the effects of the infodemic and promote more reflective and critical engagement with information online.

4.3 Negative emotions, repetitions and novelty

Harminformation campaigns frequently exploit human cognitive biases and emotional triggers to maximise their impact. Among the most potent emotional triggers are negative emotions such as disgust, anger and surprise. These emotions, when coupled with strategic repetition and the allure of novelty, can significantly influence user behaviour, leading to non-rational responses, increased belief in false content, and higher likelihoods of sharing misleading information. Understanding these mechanisms is crucial for developing effective countermeasures against Harminformation.

Content that evokes strong feelings of disgust or anger is more likely to capture attention and be remembered. Disgust, often triggered by content related to moral violations or threats to physical well-being, can create a sense of urgency and repulsion, prompting users to react quickly and share the information as a form of warning or moral outrage. Anger, on the other hand, is typically provoked by perceived injustices or threats to personal or group identity. It motivates users to take immediate action, often by disseminating the information to mobilise others or to express their outrage. Studies have shown that emotionally charged content, particularly that which induces anger and disgust, spreads more rapidly and widely on social media platforms than neutral content (Brady et al., 2017).

Repetition is another critical factor in the effectiveness of harmful information. The illusory truth effect, first described by Wired website in 2017, is a term representing a cognitive bias where repeated statements are more likely to be perceived as true. When users encounter the same misleading information multiple times, it becomes more familiar and easier to process, leading to increased acceptance. This phenomenon is particularly potent in the digital age, where algorithms can repeatedly expose users to the same content across different platforms and contexts.. Repetition drives belief in an exponential manner, with the largest increases happening during the first few exposures (L. K. Fazio et al., 2022; Hassan & Barber, 2021); therefore it is crucial to stop Harminformation early before individuals are exposed to it multiple times.

Novelty also plays a crucial role in the dissemination of Harminformation. Humans are inherently attracted to new and unusual information, which stands out against the backdrop of everyday content. This preference for novelty can be attributed to its association with learning and survival; novel information often carries potential value or risk that requires attention. Furthermore, humans like to share novel information, this is because we gain social status by spreading this kind of information. Harminformation campaigns exploit this by framing false information in novel ways or by presenting sensational and unexpected narratives. The combination of novelty and emotional arousal creates a powerful cocktail that not only captures attention but also enhances the likelihood of sharing, as users seek to inform others about what they perceive to be important or shocking new developments (Vosoughi et al., 2018).

Emotional arousal reduces the capacity for rational analysis, leading to quicker, more intuitive judgments. Additionally, the social nature of platforms like Facebook, X and TikTok amplifies this

effect, as users are more likely to share content that elicits strong emotions, thereby contributing to its spread (Pennycook & Rand, 2018).x

Brain scientists have shown that human perception and decision-making can be easily distorted under stress and fear because the brain's predictions are determined by the neural network responsible for emotions (Barrett & Simmons 2015), which explains why anxious police officers are more likely to perceive a black citizen's cell phone as a gun (known as shooter bias). Negative emotions (e.g. anger or fear) and failure to provide timely training also increase the probability of the police shooting unarmed persons (Fridman et al. 2019). This implies that stirring negative emotions in people against their government is more effective than inspiring positive emotions toward a specific country (e.g. China trying to inspire positive emotions toward their government in Taiwanese people).

In this context, interventions that reduce the emotional impact of harmful content, limit its repetition, and provide alternative narratives that satisfy the human craving for novelty without misleading can be effective.

4.4 Harminformation diffusion patterns

Harminformation tends to spread differently compared to ordinary content, in this paragraph the characteristics of how and why it spreads are analysed.

4.4.1 Farther, faster, deeper

In past years researchers have put considerable effort in identifying the spreading pattern of harmful content and in particular of false news. The phenomenon where false information spreads faster, further, and deeper than true information has been well-documented in the scientific literature. This pattern is not limited to entirely fabricated content, but extends to misleading information such as half-truths and emotionally charged content designed to disrupt society. Harminformation, which encompasses these types of content, leverages psychological and social mechanisms to achieve widespread dissemination more effectively than accurate information.

Several studies highlight how harmful content, particularly emotionally charged and misleading information, exploits human cognitive biases to achieve rapid spread. Vosoughi, Roy, and Aral (2018) demonstrated that false news spreads significantly faster, further, and deeper than true news on Twitter. Their research revealed that false stories were 70% more likely to be retweeted than true ones, reaching more people and penetrating deeper into the social network. This is partly because false information often has novel elements that capture attention and provoke strong emotional reactions, such as surprise, disgust, or anger.

The structure of social media platforms also plays a critical role in the rapid spread of harmful content. Algorithms designed to maximise user engagement often prioritise sensational and emotionally charged content, which tends to be shared more frequently. This creates a feedback loop where harmful content is continually promoted and disseminated to a broader audience. Allcott and Gentzkow (2017) discussed how social media algorithms can inadvertently amplify the spread of harmful content by prioritising engagement metrics over content accuracy (Allcott and Gentzkow, 2017).

Harminformation's ability to spread farther, faster, and deeper has profound implications for society. It can exacerbate social divisions, spread misleading content during crises, and undermine trust in institutions. COVID-19, climate change, elections and wars are topics where harmful content spreads rapidly across social media platforms, contributing to public confusion and destabilising societies. This phenomena underscores the need for robust strategies to counter the spread of information harm, including improving media literacy, enhancing content moderation, and designing algorithms that prioritise the dissemination of accurate information.

4.4.2 Echo chambers, social groups and algorithmic filtering

Social groups are fundamental units of human interaction, consisting of individuals who share common interests, beliefs, or identities and engage in regular social interactions. In contemporary society, the formation and evolution of social groups have been significantly influenced by the advent of social media platforms, which facilitate easy and frequent communication among like-minded individuals.

The influence of groups on individuals' choices and opinions has been widely studied in the literature. One famous experiment by Solomon Asch (1956) proved the influence of groups on individuals on a simple task showing how most individuals make at least once the wrong choice when influenced by the pressure of the group decisions. This result together with the one by Helfmann et al. (2023) finding that social media influencers are incentivised to have extreme positions in order to be successful might explain the increasing polarisation of opinions in communities often guided by influencers content.

Social media have made social interactions and group pressure much more frequent giving rise to the phenomena of echo chambers. Echo chambers refer to environments where individuals are exposed predominantly to opinions and information that align with their existing beliefs, reinforcing those beliefs and often leading to greater ideological homogeneity. Within these chambers, dissenting views are minimised or excluded, thus amplifying members' pre-existing views. Echo chambers can occur naturally within social groups, but social media platforms have intensified their formation by enabling users to self-select into communities that reflect their preferences and viewpoints.

Research by Flamino et al. (2021) highlights that the desire to spend more time interacting with like-minded individuals contributes to increased stance polarisation across groups. This polarisation occurs as seemingly innocuous stances and beliefs become influential markers determining social interactions, thus generating small echo chambers characterised by opinion homogeneity. The study underscores how geographic mobility and online social networks allow individuals to self-select into social environments and affiliate with groups of their choice, further solidifying these echo chambers.

Several studies have examined the role of echo chambers in exacerbating the spread of harmful content. For instance, Cinelli et al. (2020) found that social media algorithms, designed to maximise user engagement, often promote content that aligns with users' existing beliefs, reinforcing echo chambers and contributing to the spread of misleading content. Similarly, a study by Bovet and Makse (2019) demonstrated that Harminformation spreads more rapidly within echo chambers due to the lack of critical evaluation and opposition from differing viewpoints.

Echo chambers contribute to the spread of harmful content, as opinion homogeneity within echo chambers reduces the likelihood of critical examination and opposition.

Social media platforms use algorithms to filter the information available to individuals based on engagement data to determine what content to show. These data include the numbers of clicks, shares, and comments that posts receive overall, as well as users' individual platform interaction histories (Maréchal & Biddle, 2020). Unfortunately, highly clicked and shared information generally exhibits negative emotions such as anger and outrage tends to attract engagement (Brady et al., 2020; Rathje et al., 2021), and Harminformation is often based on these characteristics (McLoughlin et al., 2021; Solovev & Pröllochs, 2022).

Echo chambers are one of the most difficult problems to solve for what concerns harmful behaviours and content. In a pioneering work Avin et al. (2024) created a mathematical model representing echo chambers and analysed the possibility of fighting it with regulation; their main result is the "impossibility result", that proves, on their model, the impossibility of existence of a general regulation function that obeys to the core values of freedom of expression and user privacy while contrasting echo chambers. In chapter five some solutions and choices related to social media platforms will be listed, these solutions might counteract in part echo chambers as they balance freedom of speech with the need to stop reinforcing harmful content.

4.5 Fake news

Fake news, which can be defined as intentionally and verifiably false news (Zhang & Ghorbani, 2020), is considered to destabilise democracies, weaken the trust that citizens have in public institutions, and have a strong influence on critical aspects of our society such as elections, the economy, and public opinion (e.g. on wars). False news spread further, faster and deeper compared to true news. False news spreaders have fewer followers, follow fewer people, are less active, are less often verified users, and have been on social media less time than real users. False news is 70 percent more likely to be shared compared to real news (Vosoughi et al. 2018).

To counteract the negative effects of the spreading of fake news, several initiatives have started to appear; one widespread approach to studying and analysing fake news is fact-checking, both automatic, when AI algorithms are involved, and manual, when humans perform it.

Behind algorithms detecting false and true news there are ethical and philosophical questions of how we define truth and falsehood. In scientific literature, truth and falsehood are primarily defined by the correspondence theory of truth and the principle of empirical falsifiability. According to correspondence theory, a statement is true if it accurately reflects reality and can be consistently verified through empirical observations and repeatable experiments (Tarski, 1944). For example, "water boils at 100 degrees Celsius at standard atmospheric pressure" is true because it corresponds to observable phenomena.

Conversely, the principle of falsifiability, as proposed by Karl Popper, posits that a statement is false if it can be empirically disproven (Popper, 1959). A hypothesis such as "all swans are white" is falsified by the observation of a single black swan. Logical consistency is also crucial; any statement that violates logical principles is inherently false (Kemeny, 1952).
Given the definitions of truth and falsehood the role of fake news detectors is to check if some content (either textual, image, video or audio) has been verified through empirical observations and repeatable experiments, otherwise marking it as false means that either is logically false or there is an example of an observation that disproves it.

Manual fact-checking websites, such as FactCheck.org and PolitiFact.com, employ professional fact-checkers to analyse and detect fake news; the fact-checker's role is to compare known facts with knowledge extracted from news and assess their authenticity. Although manual fact-checking has a critical role in contrasting fake news, it is not sufficient when analysing the huge volume of newly created information, in particular for what concerns fake news spreading on social networks.

Automatic fact checking have limitations as the number of fake vs normal news labelled datasets is limited, humans are not particularly reliable labelers and databases are mostly in English and focused on political news, and, thus, covering only a small subset of news (Zhang & Ghorbani, 2020).

An additional problem in the spread of fake news is the diffusion of AI algorithms capable of generating content such as videos and images. AI-generated content can significantly increase the spread of Harminformation and manipulated content due to several factors. First, advanced AI models can produce highly realistic and coherent text, videos and images making it difficult for users to distinguish between genuine and fabricated information (Zellers et al., 2019). This realism enables the creation of persuasive fake news and misleading narratives that can be easily disseminated through social media platforms. Second, AI can be used to generate large volumes of content quickly, overwhelming fact-checking mechanisms and enabling the rapid spread of false information. Third, AI-generated content can be continuously refined to evade detection by automated systems designed to identify false content, making it a persistent challenge for digital platforms to manage (Nguyen et al., 2021). The combination of these factors creates a potent tool for the propagation of Harminformation, undermining trust in legitimate sources of information and contributing to the broader problem of Harminformation.

Until recently the biggest research efforts have focused on detecting fake news and developing methodologies to combat this phenomena (Zhou & Zafarani, 2020). However, it's crucial to understand that real news can also be harmful. A type of harmful real news is one that provokes negative emotions in vulnerable individuals or communities. Another type is the spread of excessive negative information, even if true, on trusted institutions. Therefore, it is important to balance research efforts to address other Harminformation types (Benkler, Faris, & Roberts, 2018). Prioritising a holistic approach will help mitigate the broader impact of Harminformation on society.

4.6 Confidence and doubts

A study from Aïmeur et al. (2023) shows that the percentage of people who were confident about their ability to discern fact from fiction is ten times higher than those who were not confident about the truthfulness of what they are sharing. As a result, we can deduce the lack of human awareness about the ascent of fake news.

Confidence is not necessarily a negative trait. To distinguish among a more problematic form of confidence and a more legitimate form of confidence it is necessary to define the concept of fact and opinion based confidence:

Fact based Confidence: This type of confidence is grounded in scientific evidence, logical reasoning and opinions created based on high level education.

Opinion based Confidence: This type of confidence is based on beliefs and personal experiences.

Fact based confidence is based on verified empirical observations, which, as discussed in the previous section, is the foundation on what we define as true. Being highly confident on empirically proven observations is generally not a concern. This confidence can also be based on what experts state on a topic, in particular if the topic is complex and/or novel.

Opinion based confidence is a type of confidence based on beliefs. This type of confidence is related to individuals creating opinions on topics based on personal experiences or on the beliefs their community believes in. These individuals are generally not willing to evaluate alternative interpretations and opinions as doing so would neglect their identity.

Confirmation bias tends to be a greater issue when addressing individuals showing opinion based confidence, as there is unwillingness to accept and critically analyse interpretations other than their personal one.

Harminformation campaigns can target over confident individuals, in particular on novel topics as a belief is generally created when the first information we receive on a novel topic is repeated a sufficient amount of times to be considered as true.

Spreading doubts over fact-based statements on topics is a type of Harminformation technique. One example of such a technique is represented by Oil and Gas companies spreading the idea that global warming and cooling has happened in the past. However, what they do not state is the fact that the cooling and warming cycles tend to be extremely long in time (in a timescale of hundreds of thousands of years) compared to what we are seeing today (in a timescale of decades). This kind of Harminformation technique aims to induce doubt over confident individuals believing experts stating climate change is true and caused by human activity, thus reducing the political pressure to transition to non-fossil sources of energy.

In these cases lack of confidence or trust in experts over a topic can introduce doubts in the population, protect personal interests and introduce conflicts of opinions which reduce pressure over a form of change (e.g. do not invest in renewables, stop smoking, believe that Ukraine attacked Russia and not vice versa).

4.7 Decentralised Harminformation

Decentralised Harminformation is the idea that harmful information related to the same topic is typically not platform specific. As shown by the Code of Practice Report on Disinformation made by TrustLab (a company recognized by the European Commission whose goal is to understand and counteract harmful content), Harminformation campaigns are generally part of a broader scheme of influence that generally involves several platforms at once. By spreading harmful content in different platforms attackers obtain a diversification of the target communities, and thus increase the chances of influence and destabilisation of society. Each platform manages internally the content thus making it difficult to analyse a decentralised campaign.

By targeting a variety of social media platforms, these campaigns exploit the distinct user bases, communication styles, and algorithmic structures of each platform, creating a more resilient and

pervasive spread of harmful content. This strategic diversification increases the difficulty of detection and mitigation efforts, posing significant challenges to those attempting to counteract such campaigns.

The efficacy of decentralised Harminformation campaigns lies in their ability to tailor content to the specificities of different platforms. For instance, X's character limit encourages concise, emotionally charged messages, which are ideal for quick dissemination and viral potential. In contrast, Facebook allows for longer posts and more detailed discussions, facilitating the spread of more complex narratives. Platforms like Instagram and TikTok, which prioritise visual content, enable the creation of compelling, easily shareable Harminformation through images and short videos. By customising content to suit the strengths of each platform, Harminformation campaigns can maximise their impact and engage a broader audience.

The advantages of this approach for attackers are manifold. Firstly, it enables Harminformation to penetrate diverse social networks, reaching various demographic groups with tailored messages. This broadens the campaign's reach and increases the likelihood of influencing public opinion. Secondly, decentralisation makes it more challenging for fact-checkers and platforms to identify and remove harmful content, as the same narrative may manifest differently across platforms. Lastly, decentralised campaigns can adapt quickly to platform-specific policies and algorithm changes, maintaining their efficacy over time.

One major challenge for those attempting to counteract decentralised campaigns is the sheer volume and variety of content that needs to be monitored. With each platform having its own content moderation policies and tools, a coordinated response is difficult to achieve. Moreover, the rapid spread of Harminformation across platforms means that by the time false information is identified and addressed on one platform, it may have already propagated extensively on others.

Researchers have noted that coordinated inauthentic behaviour can shift to less monitored platforms or use coded language to avoid automated detection systems (Starbird et al., 2019). This constant evolution requires ongoing vigilance and adaptation from those countering harmful content. Thus confirming the need for an Agile conceptual framework, such as the one proposed in this thesis, to coordinate stakeholders.

Furthermore, the decentralised nature of these campaigns often involves leveraging bots and coordinated networks of accounts, which can rapidly amplify harmful messages and create the illusion of widespread consensus (Ferrara et al., 2016). These tactics not only enhance the reach of information harm, but also make it more challenging to trace and attribute the sources of such campaigns, complicating efforts to hold perpetrators accountable.

5. Definition phase: Countering Harminformation

Although several existing strategies have been proposed and implemented in the scientific literature, the fight against harmful content is still in its infancy. Institutions are still trying to understand the most effective solutions and sometimes fall into the trap of considering fact checking the main one; however, in the previous chapter the listed problems show how psychology and cognition have a crucial role in this field, thus a variety of solutions are necessary.

In this chapter, for each problem identified in chapter 4, a solution is proposed and analysed. All of them are based on state-of-the-art research and on why users create, spread and/or believe Harminformation. Information harm variety requires a specialised set of tools in order to be counteracted; this is because it is a problem that, given its diversity, needs a variety of solutions.

5.1 Vulnerable communities and preventive solutions

False, misleading and emotionally charged information has different effects depending on the individuals and communities consuming it. Some communities are vulnerable to believing in non-factual information and confirming pre-existing beliefs and stereotypes. This section analyses why these communities are more vulnerable, why they are often targeted in Harminformation campaigns and what solutions can be effective in reducing the influence of harmful content on them.

Furthermore, preventive solutions are described to reduce the effects of Harminformation on social media. These solutions have shown promising results and have been shortly described in chapter 2. In this section, variants of these solutions are proposed with the aim of increasing their effectiveness.

5.1.1 Vulnerable communities

Vulnerable communities are typically characterised by several factors that limit their ability to access reliable information, make informed decisions, and defend against various forms of exploitation. Socioeconomic status plays a crucial role; lower income levels often correlate with reduced access to quality education and information resources, which hinders the ability to critically evaluate information (Benkler et al., 2018). Additionally, educational attainment is also a significant factor; lower levels of education can restrict critical thinking skills and the capacity to discern credible sources from false information (Marwick & Lewis, 2017). Digital literacy is another critical characteristic; limited experience and skills in using digital technologies can make it difficult for individuals to navigate the online information landscape effectively (Bennett & Livingston, 2020).

Another factor that contributes to their vulnerability is social isolation. Communities with less social interaction and support may be more susceptible to Harminformation due to a lack of diverse perspectives (Guess et al., 2019). Language barriers further exacerbate this issue, as non-native speakers or those with limited proficiency in the dominant language of the information being disseminated may struggle to understand or critically evaluate content (Wu et al., 2019). Trust in institutions also plays a vital role; a lack of trust in government, media, and other institutions can make communities more receptive to alternative, often misleading, sources of information (van Prooijen & van Vugt, 2018).

Psychological stress, stemming from financial and social pressures, can lead to heightened emotional responses, making individuals more susceptible to fear-based or emotionally charged information

(Guess et al., 2019). Social and cultural factors, such as strong in-group affiliations, can lead to the spread of information within echo chambers where misleading and false information goes unchallenged (Benkler et al., 2018).

Harminformation campaigns often target vulnerable communities because they are easier to manipulate. Individuals who lack critical thinking skills or digital literacy are more susceptible to being deceived (Bennett & Livingston, 2020). Furthermore, harmful information can spread rapidly within these communities, especially if it resonates with pre-existing beliefs or fears, creating an amplification effect (Guess et al., 2019). Exploiting existing distrust in authorities and mainstream media, false and misleading information offers seemingly credible alternative explanations or narratives (van Prooijen & van Vugt, 2018).

The consequences for vulnerable communities are significant. There is often a further erosion of trust in legitimate institutions and media, aggravating the initial problem (Benkler et al., 2018). Public health risks are also a major concern; the spread of false information about health can lead to harmful behaviours, such as vaccine hesitancy (Wu et al., 2019). Social division is another critical consequence, as Harminformation can deepen societal divides, fostering suspicion and hostility among different groups (Guess et al., 2019).

To address these challenges, AI techniques can be employed to identify vulnerable communities on social media and mitigate the spread of harmful content. These algorithms can analyse patterns in social media usage, socio-demographic data, interaction behaviours and education levels based on written language quality to pinpoint communities that are at higher risk (Ferrara, 2017).

AI algorithms identifying vulnerable communities should be executed by social media companies. Once identified the vulnerable communities recommendation systems should be fine tuned to include less emotionally charged, negative and misleading content addressing them. Vulnerable communities tend to have a harder life compared to ones that do not represent social minorities or do not have financial issues. Representing an overly negative digital version of life might worsen their trust in institutions and in general in society.

Although education and media literacy might be part of the solution to reduce the risk of harm, the low education level of some of these communities suggests also the possibility of lack of interest into participating to media literacy programs as other life priorities might need to be addressed first and as there might be lack of critical thinking abilities (due to low education) useful to understand media literacy programs. Therefore, there is the necessity to increase the level of content moderation for these communities.

The goal of the proposed solution is to balance the spread of negative content over communities that are already facing considerable challenges in life. In this case the recommendation system should not only consider the veracity of the information but focus on the emotional response it triggers. If the content has already been repeated and shown in the news feed of an individual it should not be recommended, moreover if an excess of negative content is shown to an individual there should be more positive and inspiring content balancing it.

In some cases there is the presence of single vulnerable individuals, generally socially isolated, which might require this solution to balance positive and negative content. Socially isolated individuals tend to be part of social media groups reinforcing their beliefs; however, theories can spread simply due to faulty information flows within these virtual communities. That is, if someone is only exposed to

information dominated by conspiracy theories, they will be more likely to believe in them (Marwick & Lewis, 2017).

Finally, these measures might be augmented in situations in time where traditional media and experts should spread critical news instead of social media. These situations include: politics, wars, new health threats, financial crisis, climate change and racial injustice. The aim is to balance news and to spur vulnerable individuals to proactively search for information on these problems rather than receiving it passively.

5.1.2 Media literacy

As discussed in chapter 4 Media Literacy interventions are provided through formal education or courses. The efficacy of media literacy has been tested in the literature; a research from Dame Adjin-Tettey (2022) found it to be effective, while, one from Guess et al. (2020) analysed the largest media literacy campaign in the world, which provided tips on how to spot false news. These found that exposure to tips improved discernment between mainstream and false news headlines. This early result supports the need to continue experimenting the adoption of media literacy as a long term solution against Harminformation. However, a work analysing the most effective types of literacy found that information literacy is most effective and the others (media, news and digital literacy) to be less effective; information literacy is defined as the intellectual framework for understanding, finding, evaluating, and using information (Jones-Jang et al. 2022).

Although more research is needed to analyse the effectiveness of literacy interventions, early results suggest that some techniques, such as adding tips on how to spot fake news and information literacy programs could be successful in increasing the capability of spotting fake news.

Literacy interventions have the limitation of being expensive. To address this issue specialised courses could be supported and incentivised in schools and workplaces nearby identified vulnerable communities, such as financially poor neighbourhoods, or places where the number of people following formal and high level education is below average. Localising interventions by targeting vulnerable communities would greatly reduce the cost of these interventions.

5.1.3 Inoculation and forewarning

Inoculation involves forewarning and preemptive corrections, which can be fact-based or logic-based, helping citizens identify misleading arguments and techniques. Large-scale field experiments, such as those conducted on YouTube by Roozenbeek et al. (2022), have demonstrated that brief inoculation interventions can enhance people's ability to recognize low-quality information. Their study highlights how manipulation techniques, such as logical fallacies and emotionally manipulative language, can be analysed and used for inoculation without prior knowledge of specific misleading content. This approach potentially provides broad resilience against social media or news content that employs these techniques.

Their work is particularly relevant as their videos explaining manipulation techniques start with an example of manipulation and then explain why the content is manipulative. This strategy has been proven to enhance people's capability to recognize malicious content. Inoculation is a cost-effective measure against certain types of harmful content and should be included as a solution to counteract

Harminformation. One potential large-scale deployment could involve regulations requiring social media platforms to show these videos during unstable and temporary periods, such as elections, new health threats (e.g., COVID-19, Ebola), or before events requiring immediate attention to factual evidence (e.g., the COP climate change conference). These inoculation videos could be translated into different languages, which would be crucial for events like the European elections, where populations speak different languages.

Moreover, Roozenbeek et al. (2022) have proven the efficacy of inoculation through eye-catching, easy-to-understand video cartoons. Therefore, regulations mandating social media platforms to use inoculation should specify the type of successful inoculation technique to be used unless other methods are also proven effective in the literature.

5.1.3 Sharing is caring

In section 2.2.1 nudging has been described as the process of asking people to focus on their behaviour while doing an action online. To combat the spread of Harminformation on social media, it is essential to implement nudges that alert users when they are about to share content from sources of questionable credibility. These nudges should be tiered, increasing in intensity with the decreasing credibility of the source. For instance, content from highly credible sources would not trigger any sharing nudges, allowing seamless dissemination of reliable information. Conversely, sharing content from less credible sources could prompt a user with a two-step confirmation process. This process would include messages warning the user about the untrustworthy nature of the content or the dubious credibility of the source. The goal is to make users pause and reconsider before spreading potentially harmful information.

Additionally, similar mechanisms should be applied when users share posts that include links they have not visited or have visited so quickly that it is unlikely they have had time to read and understand the content. In these cases, the system should flag the rapid sharing behaviour and prompt a warning or a confirmation step to ensure the user has actually reviewed the linked material. This approach leverages the concept of digital friction to slow down the impulsive sharing of unverified content, encouraging more thoughtful and informed engagement on social media platforms. Implementing such nudges could significantly reduce the spread of harmful content by prompting users to critically evaluate the content they share.

Nudges are effective on users that do not intentionally share harmful content. Intentional spreaders of Harminformation would, most likely, share the content independently from the presence of a confirmation step.

5.1.4 Confidence and doubt-spreading

Opinion based confidence is a type of confidence based on beliefs. This type of confidence is related to individuals creating opinions on topics based on personal experiences or on the beliefs their community believes in.

Being part of a community (both real and virtual) can have an impact on a person’s susceptibility to misleading information. Nowadays on social media the sharing of knowledge creates a false sense of understanding of the world. Our confidence in evaluating evidence we received from someone else can be damaging if that evidence is not shared by an expert. Humans have a limited ability to

understand the world and yet some individuals are confident in their opinions even without actually knowing the topic they are speaking about or being an expert of the field. Being in a community with shared knowledge based on this false sense of confidence can be highly damaging for the collective and individual knowledge. Therefore, it is necessary to train AI models identifying the expression of excess of confidence in social media groups and individuals in order to identify communities with a false sense of understanding. There is scarcity of these models, although few works have tackled the problem (Emerson et al. 2022, Smith et al. 2018). Most experts express detailed and precise information in a respectful non-confident way; paradoxically, this apparent lack of confidence might reduce the capability of a non-expert individual to believe a piece of factual information. An additional way to tackle this issue is to educate people to be intellectually humble and be open to the possibility of being wrong (Sgambati & Ayduk, 2023). Social media companies could create rankings of communities showing factually unjustified excess of confidence (that is confidence not backed up by facts and state of the art research) and reduce the spread of information where the sources are these communities and not field experts. Therefore, it is important for recommendation algorithms to reduce the visibility of sensitive content when over confidentially shared by non-credible sources or when identified as non factual. These interventions reduce the visibility of influencers spreading extreme and non-fact based ideologies.

A way to deprive knowledge from people is to erode their confidence. Creating doubts in a person that is confident in knowing a topic might increase the chances that person will doubt that knowledge and that confidence. Usually this strategy is used by malicious actors in order to achieve an objective. For example a person might be confident that Electric Vehicles are more environmentally friendly compared to gasoline ones, but an oil company might spread content in social media discussing that the production of Electric Vehicles is more polluting that the one of normal ones, hiding the information that on average, by including the life usage EVs are much less polluting. Although this strategy is generally used by attackers it could also be used to spread doubts over false and misleading information; in other words it could be a form of prevention against future risks, aiming to spread doubts in communities where individuals are confident about ideas that are harmful to society. For example it could be used to spread doubt on communities of people following videos having logical and false arguments about climate change.

These interventions should be experimented before being deployed at scale. In this work they are considered as future experimentations rather than ready to be implemented solutions, as few literature analyses the problem of over confidence and potential side effects of moderating content recommendation based on confidence should be analysed.

5.1.5 Community moderators

The fact that a few commenters can sway readers' opinion and can set the tone of discourse is of growing concern to internet news services and media providers. In response, some sites have introduced strict moderation of comments, as for example TheConversation.com which employs a "community manager" and has entertained options such as a "community council" to provide moderation (Lewandowsky et al. 2012).

Effective moderation within social media communities is essential to maintaining respectful and constructive discourse. Basic rules must be established within these communities to ensure that discussions remain respectful and do not incite violence, use hate speech, or target individuals or discriminated groups. Community moderators play a crucial role in enforcing these rules. They are

responsible for monitoring interactions and removing users and content that violate these guidelines. This helps to create a safer and more inclusive environment where diverse perspectives can be shared without fear of harassment or abuse.

In addition to moderator intervention, automated systems can assist in maintaining respectful communication. Algorithms can identify and flag content (including posts and comments) that are accessible to everyone and deemed unrespectful. When such content is detected, the system could prompt the content creator with suggestions for improvement. For example, the algorithm might ask the user to rewrite their message in a more respectful manner or suggest avoiding excessive use of swear words, especially when directed at strangers or individuals representing organisations. This not only promotes a culture of respect but also educates users on how to communicate more effectively and considerately.

By implementing these strategies, social media platforms can better manage their communities, ensuring that they remain spaces for positive and productive interaction. This dual approach helps mitigate the spread of harmful content and fosters an environment where users can engage in meaningful and respectful dialogue.

5.2 Incentivising deep thinking

As discussed in previous sections the problem of consuming a large amount of content in a short amount of time is that users fail to evaluate the accuracy and the credibility of the source and the content, thus are more vulnerable to Harminformation and tend to think with the fast emotional brain rather than the slow rational one. In this paragraph three solutions are explored as proposals to counteract this phenomena, these ideas are based on what psychologists and cognitive scientists have studied on users behaviour online, and thus are worth exploring.

5.2.1 Limiting information consumption

The consequences of an infodemic surpass what can be imagined. Research has found that an excess of content on social media is associated with an increased probability of depression, anxiety, lack of productivity, and low focus (Keles et al. 2020; Vannucci & Ohannessian, 2019). Given these impacts, it becomes imperative to regulate social media usage, particularly by limiting the amount and speed of content an everyday user can see each day.

This limitation, grounded in the goal of protecting mental health, could also counter harmful information. With less content presented at a slower pace, users can focus more on what they consume, increasing the chances of deeper, more rational engagement rather than shallow, emotional reactions (Twenge & Campbell, 2019). The sheer volume of information on social media often overwhelms users, making it difficult to validate the accuracy of each piece of content (Aichner et al. 2021). This cognitive overload results in fatigue, causing users to neglect details and the credibility of sources, as well as the truthfulness and potentially manipulative nature of the information they encounter (Bontcheva & Posetti, 2020). In such an environment, even high levels of media literacy may be insufficient if users, as individuals, are tired because of overstimulation and unable to engage in deep thinking.

Overstimulation is particularly problematic for younger generations, who are more likely to interact with mobile devices and multiple social media platforms simultaneously (Montag & Diefenbach,

2020). To address this issue, regulating the amount of content a user consumes, both within a specific time interval and cumulatively throughout the day, could be a viable solution. Stricter regulations might be necessary when it comes to protecting young individuals (Obar & Wildman, 2015).

Regulatory measures could include controlling the speed of vertical and horizontal scrolling and automatic content renewal, encouraging users to consume content more responsibly (Nemeth & Bonneau, 2018). Alternatively, limits could be placed on the amount of content available within a given time period. Such measures may initially seem contrary to democratic principles, but it is essential to recognize that social media, in some cases, has proven to be addictive and harmful to health (Alter, 2017). Like any powerful, addictive phenomenon, it warrants regulation. Using a social media platform could be compared to driving a car: an individual can drive freely if certain simple rules are respected (Vaidhyanathan, 2018), sometimes rules are followed because of the possibility of receiving a fine as a form of disincentive, other times physical obstacles (such as speed bumps) are added to the road to limit the car's velocity.

Although the primary motivation for regulation is mental health, a potential secondary benefit could be the reduction in the spread and exposure to harmful information and a lower amount of cognitive load, causing users to spend more time on each post and, therefore, think deeper about the information it contains. Users would see less content and thus have more time to focus on the accuracy of content, promoting deeper reflection on what they are interpreting (Bastos & Mercea, 2018).

As a car bump forces a vehicle to reduce its velocity a software could control the velocity of content consumption and limit it. In the world of social media velocity of consumption of content can be represented in two ways: 1) As the average amount of megabytes consumed per second per type of content; 2) As average number of different pieces of content (generally posts) visualised in a given minute. The first strategy refers to the megabytes consumed per second per type of content, where type of content can be video/audio/image or text, each type of content must have its own average measure as a single high definition short video can consume the amount of megabytes of an large amount of text; therefore the metric should be based also on the type of content and not only on megabytes as an absolute measure.

To ensure the effectiveness of this solution, it should first be tested on a sample of users before broader implementation (Fogg, 2003). Such a solution holds promise for improving the mental health of young generations and, concurrently, reducing the spread of harmful information.

5.2.2 Social Media Insurance

This chapter explores the feasibility and implications of implementing an insurance system to curb the spread of harmful information on social media platforms. The proposed system operates on the principle that social media users should pay a low annual fee to an insurance company, which increases if they disseminate harmful content. This financial model incentivizes responsible behaviour by making it more costly to engage in the spread of Harminformation. By extending this system to platforms like Google News, where website owners would also pay insurance premiums, the model aims to create a broader impact on information dissemination (Bak-Coleman et al., 2021; Bernal, 2021).

Central to the implementation of this insurance system is the legal recognition of social media users. To ensure accountability, users must be linked to a working ID card or official document. Social media

platforms would need to establish mechanisms for validating these IDs and communicating the information to an institution responsible for checking the document's validity. This step is crucial in creating a reliable and enforceable system where each user can be held accountable for their online actions (Gorwa et al. 2020). One possible alternative is the possibility to login into a social media platform with a digital ID card or with a service recognised by the hosting country, such as the European ID (e-ID).

The financial penalty for spreading harmful information serves as a deterrent. By increasing insurance costs, users are likely to become more conscientious about the content they share. This model encourages users to double-check the accuracy and impact of their posts, thereby reducing the likelihood of sharing harmful information. The higher cost associated with intentional dissemination of harmful content also acts as a significant deterrent, promoting a culture of caution and responsibility online (Roozenbeek & van der Linden, 2019; Vosoughi et al. 2018).

Implementing the proposed insurance system would require collaboration between social media platforms, government institutions, and insurance providers. Social media platforms would be responsible for collecting and validating user IDs, while an independent institution would verify the authenticity of these documents. Insurance providers would then adjust premiums based on the users' online behaviour and the assessments of the impact of harmful content (Gorwa et al., 2020).

An essential aspect of this insurance system is the assessment of the impact caused by harmful content. This assessment involves analysing the contribution of such content to real-life violent or extreme behaviours. For example, content that persuades or incites violence or hate speech toward social minorities would incur higher penalties. The system would need to evaluate the persuasiveness and intentionality behind the creation and spread of harmful content, ensuring that those who deliberately cause harm face higher insurance costs (Modha et al., 2021).

Penalties should use a metric capable of describing the level of harm a piece of content is doing. This is necessary as, in cases of low harmful content, such as proven unintentional false information that is not viral and did not cause negative emotions among comments the penalty should be considerably low, if not absent; while, in cases such as direct physical threat, hate speech, violence incitement or revenge porn the penalty should be extremely high. Furthermore, the penalty could be higher if the persuasiveness of such harmful content is higher; thus the need for automatic persuasive detection might be needed. A metric or system that spots extremely harmful content could be used by insurance and social media companies to flag to national authorities dangerous individuals with violent attitudes or extremely hateful toward one or more discriminated social groups. In later sections of this work, a metric addressing this issue is proposed.

5.2.3 Rewarding manual Harminformation detection

This section describes a novel approach to incentivize users to engage in deep rational thinking when consuming content on social media, rather than relying on heuristic or automatic processing often referred to as "slow brain" thinking. Social media algorithms currently create personalised content feeds by integrating user preferences into recommendation systems, thereby shaping the contextual representation of content for each user (Bak-Coleman et al., 2021; Zeng et al., 2020).

The proposed solution involves rewarding users for correctly identifying and reporting harmful information (Harminformation) on their social media feeds. This incentivization encourages users to

critically evaluate content, enhancing their ability to discern harmful information. Rewards would be small and given only if the harmful nature of the reported content is confirmed, ensuring that the system remains accurate and reliable (Vosoughi, Roy, & Aral, 2018). The incentive could be, for example, a diminished price for the insurance system described in the previous section, or a direct payment provided by the insurance company covered by social media users paying more insurance because of high harmful behaviours; another possibility would be to give for the government to provide incentives for this flagging behaviour whenever the content identified is confirmed as harmful. This kind of system could introduce in social media platforms the role of self-regulators, which is generally key to ensure safer, more respectful and fact based debates over hateful, unrespectful, misleading ones.

By motivating users to think deeply and become more informed about what constitutes harmful content, the system fosters better recognition of factual information. Additionally, the data generated from these user reports could be invaluable for creating new datasets to train AI algorithms. These datasets would help improve the accuracy of automated systems in identifying and classifying harmful information, potentially easing the burden on human fact-checkers.

One last advantage is that many social media platforms create a contextualised presentation of content to each single user based on his/her preferences. Involving a user to identify harmful content in this personalised context might work better than general purpose systems of algorithms excluding the social media experience tailored for each single user.

5.3 Harminformation regulation and detection

The proliferation of Harminformation on social media platforms poses a significant threat to societal well-being and democratic processes. Legal frameworks play a crucial role in this field by establishing standards and accountability for content dissemination. Effective regulation can help mitigate the impact of harmful content while preserving the integrity of free speech and democratic engagement.

Regulation must be implemented with caution to avoid inadvertently suppressing democratic discourse. Harmful content, such as posts evoking strong emotions such as disgust, surprise, fear, or anger should be carefully managed. Instead of outright censorship, these posts should be paused or their spread reduced, particularly on platforms that rely on algorithms to amplify content. AI algorithms should serve as the first line of defence, assessing the potential harm of the content, followed by evaluation from human experts to determine the appropriate level of visibility. This dual-layered approach ensures that content causing intense emotional reactions is scrutinised to prevent the spread of harmful information without unjustly stifling legitimate discourse.

5.3.1 Preventive and early detection

The role of content detection systems in regulating social media is multifaceted, involving both preventive and early detection mechanisms to combat the spread of harmful content. These systems are crucial in identifying and managing content such as hate speech, pornographic material, violent imagery and incitement to violence that can pose significant risks to public safety.

Preventive detection refers to the use of AI algorithms to identify and block harmful content before it is published. This proactive approach is designed to prevent the dissemination of dangerous material

at the source, thus reducing the potential for harm. AI-driven preventive detection systems scan text, images, and videos for indicators of prohibited content based on predefined criteria. This process is essential for maintaining a safe online environment by filtering out explicit or dangerous content in real-time. However, the complexity of human language and context means that these systems are not infallible and can sometimes miss subtle or emerging forms of harmful content.

Early detection, on the other hand, deals with content that has bypassed preventive filters or has become harmful in a broader context. This includes content that may appear benign initially but becomes problematic when it gains traction or is manipulated to target specific communities. Early detection systems are tasked with monitoring published content for signs of virality and context-dependent harm. For instance, content that incites disgust, surprise, or anger in user comments can be flagged for review. Such reactions can be indicative of the content's potential to incite violence or spread Harminformation. Cavaliere et al. (2023) propose an example of an early detection system based on the virality and negative emotionality of topics on twitter. These early detection systems can contribute in the fight against Harminformation.

These systems are expensive when checking videos, as the amount of data to scan is considerable. One approach to this problem could be checking only videos from users and people with a high amount of followers or producing content on average going viral. Additionally, for long videos a good strategy could be to analyse the first minute and then analyse only the text obtained as automatically generated subtitles; or alternatively to sample the images in the video, with higher sampling rates in the first minutes of the video. This is useful as the first minute of a video is the one that could catch users attention, and text can be analysed by AI algorithms much more efficiently.

The integration of both preventive and early detection systems is essential for a comprehensive approach to content regulation. Preventive detection serves as the first line of defence, while early detection systems provide a secondary layer of scrutiny to catch harmful content that slips through the initial filters. This dual approach helps to address the dynamic nature of online content, where the context and impact can evolve rapidly.

5.3.2 Regulating content by law

In recent times, co-regulation and direct regulation have begun to take hold in several countries, primarily targeting companies that own large social media platforms. These regulations are designed to reduce the spread of harmful content. By mandating that these companies invest in mechanisms to curb the dissemination of harmful content and transparently demonstrate their efforts, these regulations tend to be effective as companies are sanctioned if not respected.

A study from Kozyreva et al. (2022) asked US citizens whether they would remove problematic social media posts and whether they would take punitive action against the accounts. Respondents were shown key information about the user and their post. The majority preferred removing harmful content over protecting free speech. Respondents were more reluctant to suspend accounts than to remove posts and more likely to do either if the harmful consequences of the content were severe or if sharing it was a repeated offence. This survey shows how citizens understand and accept the need for content moderation, thus justifying laws regulating content online.

Recent legislation, such as the EU Digital Services Act (DSA) and the Online Safety Act in the UK (Department for Science, Innovation & Technology, 2024), compels private companies that manage

social media and news platforms to manage harmful content to mitigate the risks of an internal crisis. Companies must now prioritise reducing the risks associated with a Harminformation crisis and its potential financial and legal repercussions. As Coombs (2014) states, crisis management protects an organisation not only from the immediate harm of a crisis but also from secondary harm resulting from legal action.

In December 2020, the European Commission initiated efforts to ensure that social media platforms moderate harmful content upon detection. The DSA aims to create a co-regulatory framework with voluntary Codes of Conduct to address systemic risks, including those related to harmful content (Casero-Ripollés et al., 2023). While the Commission pushes digital platforms to combat the spread of Harminformation, industry leaders argue that it is not their responsibility to control material that is not explicitly illegal. Social media companies have partnered with EU fact-checkers to debunk false reports, but these partnerships have had limited success in accurately targeting false narratives.

There is a tension between securitization, which treats Harminformation as a serious threat requiring strong measures, and self-regulation, which relies on voluntary actions and minimal intervention. This contrast creates dissonance, as the EU is accused of both promoting a strong stance against harmful content and being lenient with digital platform companies' obligations. This internal contradiction is something the new DSA addresses. Resolving this contrast is crucial for shaping future EU policies and determining their effectiveness in protecting European democracy (Heiskala & Aro, 2018).

Co-regulation and direct regulation have some limitations. They clearly regulate hate speech, incitement to violence, and pornography, but they often leave the responsibility to the platforms of moderating more complex information spreading patterns causing harm (such as negative emotional content related to sensitive topics or misleading content). Moreover, not all platforms adhere to direct regulations and content moderation requests. For example, X (formerly Twitter) and its owner, Elon Musk, have demonstrated non-compliance. X refused to remove violent content showing a stabbing in a Sydney church (Taylor & Butler, 2024) and has not complied with the new DSA by failing to provide an adequate ad repository and blocking access to data for researchers, leading Musk to threaten legal action against the EU.

It is essential to recognize that content eliciting strong negative emotional responses should be regulated, as current regulations often overlook this aspect. Nevertheless, emotional content can sometimes be truthful and play a positive role in raising awareness about critical issues. Therefore, when dealing with real news, the visibility of such content should be selectively limited, particularly for vulnerable individuals and communities who may be disproportionately affected. The immediate reduction in visibility of potentially harmful content allows for thorough analysis and evaluation by human experts, balancing the need to protect the public from Harminformation with the preservation of informative and necessary content. This approach is pivotal in maintaining a healthy information ecosystem where the benefits of rapid information dissemination are not overshadowed by the risks of unchecked harmful content.

5.3.1 Self-regulation

Harminformation is a complex phenomena given its variety of types, targets and application in different contexts (e.g. different platforms). Social media platforms differ in how content is spread, and therefore each platform has its own specific characteristics. In this context, it is crucial for each platform to find ways to self-regulate content that could be harmful to individuals.

The goal of self-regulation should be to voluntarily manage all types of harmful content that standard regulations fail to address. For instance, the European DSA obliges companies to self-regulate and report on their actions to reduce the spread of harmful content (European Commission, 2022). This indicates that self-regulation is not always a voluntary act, but often involves self-identifying and implementing content-regulating strategies.

Outside the European Union, where regulations and laws are less clear in some regions, social media companies often have to delineate between content regulation and free speech. For example, Instagram (and Threads) introduced a policy describing political content as "potentially related to things like laws, elections, or social topics," with social topics including issues like international relations or crime (Meta, 2023). The policy impacts the recommendation system, avoiding proactive recommendations of political content from accounts an individual does not follow.

This policy is likely to reduce issues of confirmation bias and manipulated political news. Although Meta's decisions currently affect only political news, it is conceivable that future regulations could target other types of harmful news, such as gender inequality, climate denial, harmful public health information, and sensationalistic news undermining scientific consensus or discriminating against minorities. In this regard, AI plays a significant role in identifying harmful content, enabling social media platforms to avoid recommending it.

5.4 Additional solutions

5.4.1 Identifying user polarisations and social groups

The recognition and mapping of online communities and their ideas are crucial steps in combating the spread of false information, sensationalistic content, and content that incites violence. By identifying and analysing these communities, it becomes possible to determine which groups are more likely to disseminate harmful information. This mapping process should be conducted by independent organisations, supported by institutions like the European Union, to evaluate whether the content of an online community should be restricted or blocked.

Identifying polarisation and social groups is essential in mitigating the spread of harmful information such as hate speech, pornography, racism, and other forms of harmful content. Polarised groups often have higher probabilities of sharing such content, thus understanding their dynamics can aid in preemptive measures against the dissemination of harmful information.

Monitoring groups that exhibit disrespectful, sensationalistic, and negative behaviour is imperative to reduce societal harm and prevent destabilisation on sensitive issues like COVID-19, elections, climate change, and misleading news about jobs and the economy.

5.4.2 Labelling harminformation

Labelling is the process of adding a label or correction to a piece of content. In the scientific literature it has been proven that detailed corrections might be perceived as a stronger persuasion attempt than a simple correction. While it does not threaten one’s freedom when facts are remembered, it becomes threatening when one’s attitude is about to change, especially if perceived to be forced by longer, more detailed corrections (Lewandowsky et al., 2012). This notion highlights the delicate balance required

in correcting Harminformation without triggering defensive reactions that could solidify the false/misleading beliefs further.

Corrections are rarely fully effective; despite being corrected and acknowledging the correction, people often continue to rely at least partially on information they know to be false. This phenomenon, known as the continued-influence effect, has been observed across various materials and modes of testing, persisting even when participants are warned at the outset that they may be misinformed (Ecker et al., 2010). In some cases, particularly when corrections challenge deeply held worldviews, belief in the false information may even increase.

Showing text-warnings to news in social media makes people not believe the news, but in contrast failing to label a false news as Fake, in presence of warnings, makes them look like “true” to the public. Thus the “implied truth effect” (Pennycook, 2018).

In social media, rather than merely showing labels and corrections, it might be more effective to recommend accurate information to the user, presenting it repeatedly as normal posts (e.g. posts suggested by the social media itself from reliable sources). This approach can help integrate correct information into the user’s feed naturally, reducing the resistance that might arise from overt corrections. Google search engine has a similar approach during crisis, showing as first results information from trusted sources and suggesting it.

Instead of corrective messages, letter labels similar to those used for food in supermarkets could be employed. These labels could be based on a Harminformation metric, categorising content from A (safe) to E (highly harmful). This method would provide a straightforward and non-intrusive way to inform users about the reliability of the content they encounter, potentially reducing the spread of harmful information without provoking defensive responses. In section 5.6 an Harminformation metric is defined, the proposed metric could be used for letter labelling information.

Content generated by AI, such as text content, video and/or images, must contain metadata by law that give the possibility to social media companies to distinguish between AI generated content and real one. Removing the metadata with an external tool must be considered illegal. Furthermore, social media companies and, more in general, Very Large Online Platforms (as defined by the EU DSA), must show to the user a label notifying that the content has been generated with an automatic tool using AI.

5.4.4 Free fact-based journalism

Free fact-based journalism is essential for combating Harminformation, serving as a cornerstone for informed public discourse and democracy. Trustworthy journalism provides accurate, verified information, enabling citizens to make informed decisions and fostering resilience against misleading content. In contrast, environments where professional news media are distrusted see an increased reliance on alternative sources, such as online platforms known for distributing harmful content (Tsfati & Cappella, 2003). This phenomenon occurs because people in such environments are less likely to engage with diverse sources of political information and critically evaluate them (Benkler, Faris, & Roberts, 2018). Consequently, societies with high distrust in professional news media have lower resilience to disinformation.

Knowledge plays a critical role in how individuals process information. As people become more knowledgeable about a topic, their perceptions are less influenced by confirmation bias and naive realism (Ross & Ward, 1996). Therefore, robust public service media are crucial in exposing individuals to varied political information sources, helping them critically assess and resist online disinformation. Research has demonstrated that increased knowledge enhances people's ability to navigate information, reducing susceptibility to misinformation (Prior, Sood, & Khanna, 2015).

Fact-based journalism, grounded in rigorous standards of accuracy and impartiality, is pivotal in maintaining a well-informed public. It counters the spread of false information by providing reliable narratives that can correct or mitigate the impact of misleading content. The role of journalists in verifying facts and presenting balanced views is fundamental to ensuring that the public receives a clear and accurate picture of events and issues. By doing so, journalism fosters a more informed and engaged citizenry, capable of critically evaluating information and making sound decisions.

Moreover, the presence of credible news sources helps build trust in media institutions. This trust is essential because it encourages the public to turn to these sources for information rather than unreliable alternatives. When people trust the news media, they are more likely to be exposed to a variety of viewpoints and less likely to fall prey to Harminformation campaigns. Trust in media thus acts as a bulwark against the proliferation of false information, strengthening societal resilience.

In sum, free fact-based journalism is indispensable in the fight against harmful information. It provides the public with reliable information, fosters critical thinking, and builds trust in media institutions. Ensuring the robustness of fact-based journalism is crucial for maintaining an informed public and a healthy democracy.

5.5 APIs to counteract distributed Harminformation

In the context of harmful content distributed across multiple online platforms, the use of Application Programming Interfaces (APIs) to insert references to posts and their content into a centralised database managed by an institution could be a crucial measure. APIs are sets of protocols and tools that allow different software systems to communicate with each other. In this case, APIs interfaces provided by institutions (e.g. the EU commission) and utilised by accredited social media companies would allow for the insertion of anonymous textual information into this central database. Simultaneously, trusted companies or governmental organisations could analyse this content in real-time. This dual approach would enable the clustering of topics and identification of emotionally charged or misleading content that is not detectable from a single source. The distribution of harmful content across multiple platforms can be a strategic move by Harminformation attackers to reach a diverse range of targets while complicating detection efforts for individual platforms. By spreading harmful content in this way, attackers exploit the fragmented nature of online content moderation, thereby increasing the challenge of timely and effective intervention.

On the opposite side, social media platforms can provide APIs interfaces to institutions that can use them to request automatic content moderation. This mechanism could significantly enhance the speed and efficiency of content moderation efforts, especially in the face of distributed Harminformation campaigns. Automated content moderation, facilitated through these APIs, would allow for the rapid identification and management of harmful content before it has the chance to proliferate extensively. This real-time analysis and response system would be particularly beneficial in counteracting the spread of harmful content that might otherwise evade detection due to its dispersed nature across

various platforms. Finally, considering the delicate nature of such services, the APIs must adhere to the strongest security protocols available, to avoid third actors intercept data, insert non existent data in the database or unauthorised actors to ask for content moderation.

An additional advantage of aggregating harmful content data via APIs into a single database is the potential for conducting higher quality research. This could provide valuable insights into the nature and spread of information harm. The EU DSA already includes provisions for accessing data for research purposes, but it does not facilitate real-time analysis. By enabling real-time data analysis through APIs, not only can immediate threats be managed more effectively, but researchers can also gain a deeper understanding of Harminformation dynamics and develop more robust strategies for mitigating its impact.

By centralising the monitoring and analysis of potentially harmful content, institutions can gain a comprehensive view of trends and tactics. This centralised approach not only streamlines the detection process but also ensures that different platforms can collaborate more effectively in combating the spread of harmful content.

5.6 Harminformation metrics

In previous sections, we have defined the problems related to Harminformation and proposed a wide variety of existing or new solutions. However, one of the most significant challenges lies in defining harmfulness itself. Natural language has limitations in describing complex phenomena, and the scientific literature often seeks clear definitions. Harmful information can be abstractly defined, but such definitions are subject to interpretation. Information can be harmful in various ways to individuals, communities, and organisations. Therefore, precise metrics are necessary to specify what makes information harmful online, and, in particular, what Harminformation stands for in the context of social media and online news.

With the goal of giving a precise and non-interpretable idea of what is Harminformation, in this thesis, we propose two metrics, primarily applicable to social media (but not exclusively). The first metric analyses the harmfulness of a single piece of information, such as a social media post. The second metric assesses the harmfulness of topics, where each topic consists of individual pieces of information (e.g., a list of social media posts and news about the same topic). The metric is obtained by computing the average scores of each individual piece of information belonging to the topic. Therefore, the second metric uses the first one.

These metrics aim to provide a clear definition of harmful information. They are based on scores obtained from algorithms that represent different types of harmful content (e.g., hate speech, falsehood, emotional charge). The factors are grounded in the scientific research cited in this work and reflect state-of-the-art results in the field.

In the following sections, we present these metrics, along with citations to algorithms needed to compute the factors. The proposed metrics are based on what is intuitively considered more harmful in scientific research. However, the weights used in the metrics can be adjusted according to the specific application needs of an application.

When a metric surpasses a threshold an expert is flagged, this behaviour has been chosen to give the possibility for humans to detect false positives. Additionally, not all harmful information should be

removed, in some cases society needs the typical characteristics of harmful information to alert people of an incoming danger (for example alerting people of an imminent hurricane), thus the need for careful and human-checked content moderation.

Flagging an expert can mean different things depending on the desired goal of an application. For example it can mean to send an API request to ask for content moderation, send an email to fact checkers of an organisation, or flag defence officials in case of highly harmful content. In any case an additional check performed by humans aimed at understanding content harmfulness is strongly advised even when the algorithm returns a high value for a metric.

The capability to compute the factors and metrics depend on the availability of data and metadata. Intuitively, considering the direct accessibility to data of social media companies, large search engines companies (e.g. bing) and news aggregators (e.g. google news), it can be imagined that these algorithms should be computed by them, while for news outlet analysis private and public organisations in strict contact with institutions should be partially responsible of analysis, given the distributed nature of them. Additionally, for what concerns aggregated topic measures, the metric results could be shared by individual companies into a centralised database, where a topic metric can be averaged over platforms.

5.6.1 Single content harmfulness

In this section, we present an algorithm designed to compute a metric for assessing the harmfulness of individual pieces of content. The algorithm evaluates various types of harmful information and applies corresponding weights to derive an overall harmfulness score. For certain types of harmful content, such as hate speech and incitement to violence, the algorithm flags the content for expert review immediately when a specific threshold is reached based on a single factor score. For other types, the flagging occurs when a weighted sum of the factors surpasses a predefined limit, with each factor representing the normalised score of a type of harmful information.

The adopted threshold for flagging content is set at 70, though this can be adjusted depending on the application's requirements. For instance, a threshold of 90 might be used for detecting only extremely harmful content. The weights assigned to each type of harmful information are based on what is considered more harmful according to the scientific literature.

This algorithm ensures a comprehensive assessment by considering multiple types of harmfulness, resulting in a balanced and well-rounded score. The Harminformation types are grounded in research and scientific results previously cited in this thesis. The factors used in the algorithm correspond to the variables listed in Table 1.

To calculate the harmfulness metric, each factor is normalised to a scale from 0 to 1 to standardise the inputs. A weight is then assigned to each factor based on its perceived level of harm. The weighted sum of all normalised factors is calculated and scaled to a range of 0 to 100. Content is flagged for expert review if the weighted sum reaches 70 or above. Alternatively, if the score of single highly harmful information factors (that are hate speech, incitement to violence, violent content, and pornographic content) exceeds 50, the content is flagged for expert review.

The pseudocode below, shown in Figure 6, illustrates the detailed steps to calculate the single content harmfulness metric, including the thresholds and weights assigned to each factor. Weights are based on the level of harm done a characteristic of the content can do. This level of harm is assigned

intuitively based on common-sense and scientific results. For example, it is clear that hate speech and incitement to violence can be more harmful compared to high confidence text (when analysing a single piece of content).

```
def calculateHarmfulnessScoreForContent(
    emotionalNegativity, vulnerableCommunityScore, sensationalism, overConfidence, novelty, repetition, falseness, hateSpeech, incitementToViolence,
    manipulativeFraming, sourceIncredibility, intentionToMislead, violentContent, pornographicContent
):
    # Normalize each factor to a scale of 0 to 1
    norm = normalize(
        emotionalNegativity, vulnerableCommunityScore, sensationalism, overConfidence, novelty, repetition, falseness, hateSpeech, incitementToViolence,
        manipulativeFraming, sourceIncredibility, intentionToMislead, violentContent, pornographicContent
    )

    # Assign weights to each factor based on importance
    weightEmotionalNegativity = 0.08
    weigthVulnerableCommunityScore = 0.05
    weightSensationalism = 0.07
    weightoOverConfidence = 0.04
    weightNovelty = 0.05
    weightRepetition = 0.05
    weightFalseness = 0.09
    weightHateSpeech = 0.11
    weightIncitementToViolence = 0.11
    weightManipulativeFraming = 0.05
    weightSourceIncredibility = 0.07
    weightIntentionToMislead = 0.06
    weightViolentContent = 0.09
    weightPornographicContent = 0.08

    # Calculate the total harmfulness score
    harmfulnessScore = (
        norm.emotionalNegativity * weightEmotionalNegativity +
        norm.vulnerableCommunityScore * weigthVulnerableCommunityScore +
        norm.sensationalism * weightSensationalism +
        norm.overConfidence * weightoOverConfidence +
        norm.novelty * weightNovelty +
        norm.repetition * weightRepetition +
        norm.falseness * weightFalseness +
        norm.hateSpeech * weightHateSpeech +
        norm.incitementToViolence * weightIncitementToViolence +
        norm.manipulativeFraming * weightManipulativeFraming +
        norm.sourceIncredibility * weightSourceIncredibility +
        norm.intentionToMislead * weightIntentionToMislead +
        norm.violentContent * weightViolentContent +
        norm.pornographicContent * weightPornographicContent
    ) * 100

    # Immediate review thresholds for highly harmful factors
    highHarmThreshold = 0.50

    if norm.hateSpeech >= highHarmThreshold
       or norm.incitementToViolence >= highHarmThreshold
       or norm.violentContent >= highHarmThreshold
       or norm.pornographicContent >= highHarmThreshold:
        flagForExpertReview()
    elif harmfulnessScore >= 70:
        flagForExpertReview()
```

Fig. 6 Pseudocode for the calculation of the single content metric

```
def calculateHarmfulnessScoreForTopic(contentList, virality):
    totalScore = 0
    contentCount = len(contentList)

    for content in contentList:
        totalScore += calculateHarmfulnessScoreForContent(
            content.emotionalNegativity,
            content.vulnerableCommunityScore,
            content.sensationalism,
            content.overConfidence,
            content.novelty,
            content.repetition,
            content.falseness,
            content.hateSpeech,
            content.incitementToViolence,
            content.manipulativeFraming,
            content.sourceIncredibility,
            content.intentionToMislead,
            content.violentContent,
            content.pornographicContent,
        )

    # Average score for the topic
    avgContentScore = totalScore / contentCount

    # Normalize virality
    normVirality = normalize(virality)

    # Add weight for virality
    weightVirality = 0.10

    # Final topic harmfulness score
    topicHarmfulnessScore = (avgContentScore + (normVirality * weightVirality)) / (1 + weightVirality)

    return topicHarmfulnessScore

# Example usage
topicHarmfulnessScore = calculateHarmfulnessScoreForTopic(contentList, virality)


def topicClustering(allListOfContents):
    # here use a topic clustering tencique
    return clustersOftopics


def computeHarminformationMetrics(allContentList):
    clustersOftopics = topicClustering(allContentList)

    for contentList in clustersOftopics:
        topicHarmfulnessScore = calculateHarmfulnessScoreForTopic(contentList)
        if (topicHarmfulnessScore > 0.70):
            flagForExpertReview()
```

Fig. 7 Pseudocode for the calculation of the aggregated topic content metric

5.6.2 Topic harmfulness

Humans disseminate concepts by describing them in various ways while trying to maintain the core meaning. Analysing individual pieces of content independently has the limitation of not identifying harmfulness that manifests at the topic level rather than through single instances of information. To address this issue, one effective approach is to aggregate the single content metrics within the same topic to derive a unified metric for each topic. This requires implementing a topic clustering algorithm, which assigns each piece of content to one or more clusters. Once these clusters are calculated, a metric for each topic can be computed. This is achieved by averaging the harmfulness scores of all content pieces within a cluster representing a topic.

Given the vast amount of information on social media, it is practical to prioritise fact-checking and monitoring content that exhibits a rapid diffusion rate and spreads widely and deeply. Harmful content that reaches only a few individuals typically lacks the capacity to destabilise society. Therefore, enhancing the aggregated metric with a virality factor is crucial. Recognizing whether a topic is going viral is essential for the early detection of highly harmful information related to that topic, enabling timely interventions to mitigate its spread.

By integrating the virality factor into the aggregated metric, we can better identify and respond to harmful topics that have the potential to spread rapidly and broadly, enhancing our ability to protect individuals and society from the detrimental effects of harmful information.

Since this is an aggregated metric, a threshold for the topic harmfulness score is adopted—in this algorithm, the threshold is set at 70. However, as previously discussed, the threshold can be adjusted according to the specific application.

The pseudocode in Figure 7 illustrates the calculation of the topic harmfulness metric, including the weight assigned to the newly incorporated virality factor.

5.6.3 Computing factors

In previous sections the single content and aggregated content (topic) metrics have been proposed as a formal way to detect Harminformation. The metrics give the possibility to get information on why some information is harmful and if it is harmful in the context of a topic, as a single piece of information or both.

The metrics are calculated starting from factors which are scores based on what the literature has found harmful in terms of information. Giving researchers an indication on how to calculate these scores is crucial in order to implement the metric. Therefore, in this section for each factor a scientific work computing the metric is cited. The cited works can be of inspiration to adapt the calculation of the scores to a specific application domain; thus, they provide a solid starting point to the implementation of the Harminformation metrics.

In table 2 the factors are listed together with citations of works implementing them or scientific review suggesting a wide variety of implementations.

Name	Citations
Emotional Negativity	Akinpelu et al. 2023, Kumar & Martin 2023, Ravanelli et al. 2021, Demszky et al. 2020
Vulnerable community detection	Mossie & Wang 2020, Von Luxburg 2007, Que et al. 2015
Sensationalism	Gonzalez et al. 2023, Naeem et al. 2020
Overconfidence	Emerson et al. 2022, Smith et al. 2018
Novelty	Gruhl et al. 2021, Ghosal et al. 2022
Repetition	Number of times the content has been shared. Available as metadata of a post or by counting the number of news in the same topic cluster (when not applied on social media data).
Falseness	Capuano et al. 2023
Hate Speech	Zhou et al. 2020, Modha et al. 2021, Del Vigna12 et al. 2017
Incitement to Violence	Khan et al. 2024, Saha et al. 2023
Manipulative Framing	Card et al. 2015
Source Incredibility	Wijesekara & Ganegoda 2020
Intention to Mislead	Intentionality can be measured as “proximity to the root source of the content creator”. 1 if the user is the original poster, 0.5 if the user has a close “social media " relationship with the poster, 0 if a post is shared but the user is not the original poster. Creating a content is a highlty intentional spread of information, sharing it because of close connection is also, in some way, intentional spread of information. Sharing it with no social closeness should be considered as an unintentional spread of harmful content.
Violent Content	Mohammadi et al. 2016, Nova et al. 2019, Mondal et al. 2017, Naik & Gopalakrishna 2021
Pornographic Content	Perez et al. 2017, Marcial-Basilio et al. 2011
Virality	Elmas et al. 2023, Cheng et al. 2014, Weng et al. 2013

Table 2 Citations of works calculating the factors used in the Harminformation metrics

5.6.4 Topic clustering

Clustering topics, in particular in social media, is a necessary step to compute the aggregate topic metric. Topic clusters can be computed in different ways. Wartena & Brussee, 2008 computes it without any prior knowledge of category structure; Comito et al. (2019) propose an online algorithm to discover topics that incrementally groups topics from text; Williams et al. (2016) propose a tool called GeoContext to detect topics; a final example of work tackling the issue of topic clustering is the one by Dikovitsky & Fedorov (2020) which use Word2Vec and K-means techniques to cluster the topics. Independently of the chosen method, topic clustering is crucial to find topics which have a higher Harminformation aggregate metric compared to the others. Furthermore, having average factor scores associated with each topic can be helpful to policy makers to understand what characteristic each topic expresses.

5.6.5 A visual overview of the metrics

In this section a visual representation of how the metrics are calculated starting from various inputs is shown. Considering a source of content (e.g. a social media, or an aggregate of news websites) and a timeframe going from T_1 to T_n and imaging N pieces of content of different nature (image, video, text) to be published from the source the metrics can be calculated as described in section 5.6.1 and 5.6.2. Therefore, the image below is a visual representation of how to integrate those algorithms to compute the metrics. Each type of input has different preprocessing steps:

- Text: Textual information can be converted into sentence embeddings (Yang et al. 2020).
- Image: Object detection (Ren et al. 2015, Girshick et al. 2014) and image to text (Khan et al. 2021) should be performed.
- Video: for efficiency only samples of images should be analysed. Ideally in the first minute the sampling rate should be higher, while over time it should decrease. This is because the first minute is the one attracting the attention of the user or to make them skipt the content. The audio (speech) of the video should be translated into text (Graves et al. 2006) and then into sentence embeddings; all of it should be translated if the video is short; while only the first 5 minutes can be translated if the video is long. The sampled images should be preprocessed as described above for images (that is object detection and image to text transcription).
- All types: all types of content have common information associated with them. In particular, in social media, the comments must be extrapolated, together with metadata of users writing comments and metadata of the user creating the content. Furthermore, the post metadata should be considered.

After preprocessing of inputs, preprocessed data of each content is assigned to different kinds of analysis necessary to compute the Harminformation factors. Furthermore, the preprocessed data is passed to the topic clustering algorithm, where the content will be clustered in an online fashion or added in a dataset to be clustered in batch. Highly harmful factors can directly cause expert flagging, alternatively a high weighted sum of all factors on a single input content can cause flagging. For each content the factors scores and weighted sum are passed to the belonging cluster of topics where they are added to the average of the aggregated topic metric. Finally, in case of a highly aggregated topic metric an expert is flagged. In Figure 8 the flow of data and its direction is represented as arrows; colours are used to contextualise data flows about each type of subprocess.

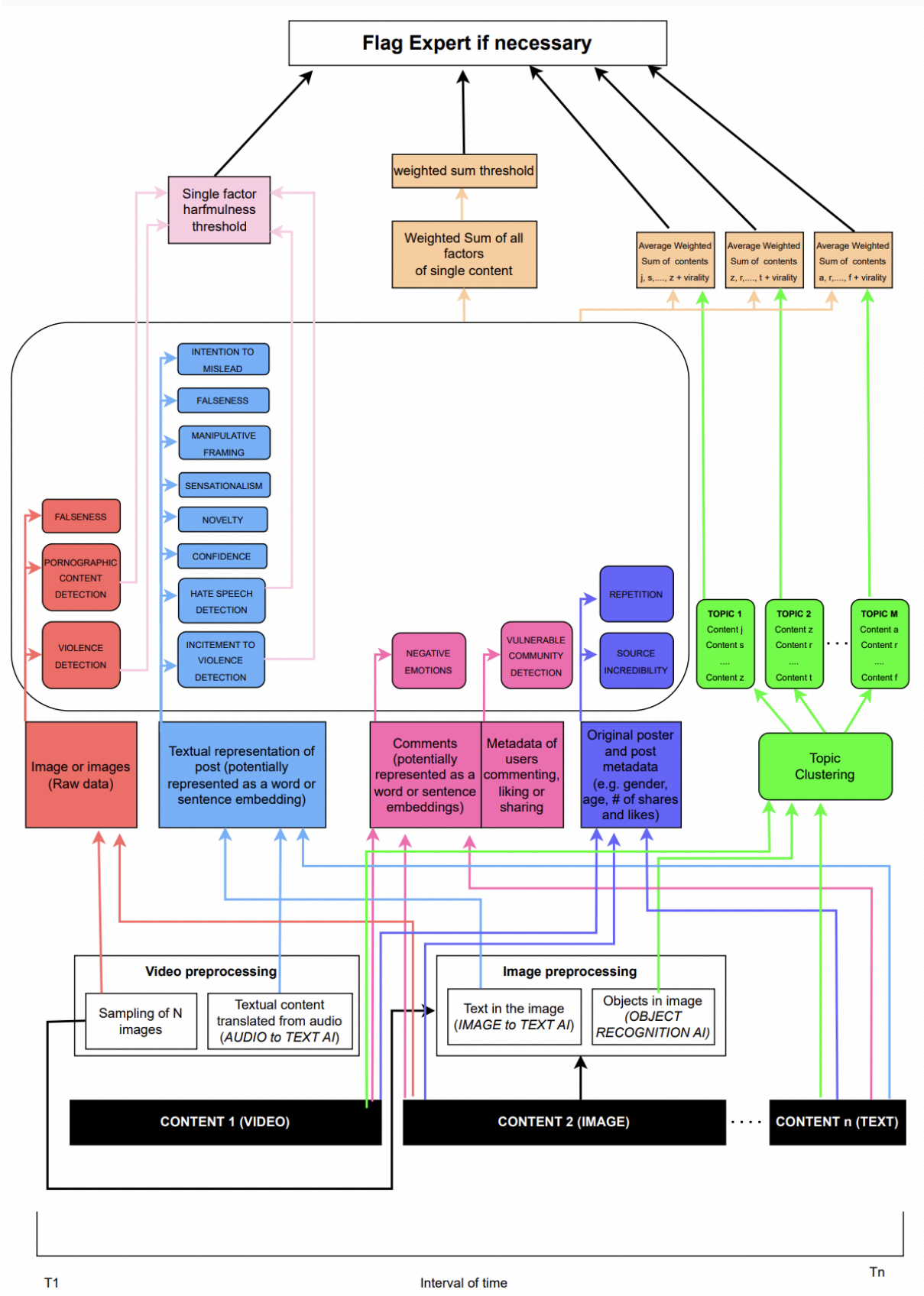

Fig. 8 Visual representation of the computation of Harminformation metrics

5.6.6 Adapting the metrics

The proposed algorithm for evaluating the harmfulness of information can be further refined and improved by incorporating additional factors. These additional factors, identified through ongoing research and experimentation, could enhance the algorithm's accuracy and reliability. Experimentation is crucial to validate the proposed weights for each factor, ensuring they accurately reflect the potential harm posed by the content.

An important improvement involves adapting the algorithm to only use factors that are computable at any given time, considering the available data and metadata. This adaptability ensures the metrics remain functional even when certain data points are missing. Additionally, the implementation should include a mechanism to verify the presence of a minimum required number or combination of factors necessary to generate a reliable harmfulness score. This ensures that the metric remains robust and meaningful, even with incomplete data.

Each normalised factor can be integrated into a graphical user interface (GUI) of a software application. This interface would visually represent the harmful characteristics of a piece of content or a topic, providing users with an intuitive understanding of the different aspects contributing to its Harminformation score. Such a GUI would be invaluable for content moderators, researchers, and policymakers in assessing and addressing harmful information. Let us note that although in the aggregated metric pseudocode the score for each factor is not computed it can be easily computed by averaging the scores of a factor for each piece of content belonging to a topic (e.g. to calculate the aggregates emotion intensity it can be averages over all posts related to the 2024 European Election).

Preventive detection of Harminformation is another critical application of the "single content" algorithm. By analysing content before it is published, the algorithm can serve as a proactive measure to identify and mitigate potential harm. Repeated and periodic analysis can detect patterns of repetition, which is another indicator of potential harm.

For the aggregated topic-level metric, it is essential to periodically run topic classification algorithms. This ensures that both new and existing topics are regularly evaluated, providing up-to-date harmfulness metrics.

Furthermore, storing all factors and metrics over time allows for the development of a time-aware metric. This temporal perspective enables the detection of evolving harmful trends in individual pieces of content and broader topics, offering a dynamic view of the information landscape.

An additional improvement could be to create a score representing the average information harm over all topics. This can be useful when it is not the single topic creating harm, but the sum of many overly negative and harmful topics.

6. Prioritisation

Prioritisation is the process of defining which solutions should be implemented first. Economic resources are limited, and some solutions are more effective and easier to implement. In this paragraph the process of prioritisation is simulated as it should be performed by a group of experts belonging to different private and public organisations.

In table 4 the solutions discussed in chapter 5 are listed, together with the expected time of implementation, outcome and cost. These three variables are then averaged to obtain a priority score. Each variable assumes values from 1 to 10 with the following meaning:

- **Velocity:** 1 for low velocity of implementation time 10 for fast implementation time;
- **Expected positive outcome:** 1 for low positive expected outcome 10 for high expected positive outcome;
- **Cost:** 1 for high cost and 10 for low cost;

Both expected positive outcome and cost are divided in short term and long term. Short term outcome is related to the positive outcomes that a solution has over time. For example, inoculation and media literacy expected positive outcomes are higher in the short term as it is proven that inoculation and media literacy positive outcomes slowly fade over time. Short and long term costs refer to the immediate costs in the short term to implement the solution and the long term costs to maintain it.

Let us note that implementation time, expected outcome and cost are based intuitively on scientific works; however, the aim of this paragraph is to be purely indicative of the prioritisation process.

Solution name	Velocity	Expected positive outcome		Cost		**Priority**
		Short term	Long term	Short term	Long term	
Inoculation	8	8	7	9	8	8
Friction when sharing from non credible sources or overly negative content (sharing nudges)	8	7	7	9	9	8
Reduce content velocity spread for almost-viral topics (reduce information overload)	7	8	9	7	7	7.6
Preventive and early detection	6	9	9	6	7	7.4
Community moderators	8	7	6	7	7	7

Reducing the exposure of vulnerable communities/individuals to overly negative content	7	7	7	7	7	7
Users pay an annual fee to an insurance company, which increases if they disseminate harmful content	6	9	8	5	5	6.6
Harminformation detection and labelling	5	7	8	5	8	6.6
Polarised group detection	7	6	6	7	6	6.4
Letter labelling information	5	6	8	5	7	6.2
Free fact-based journalism	3	8	9	3	8	6.2
Regulation to add metadata to content generated with AI	5	5	8	5	8	6.2
Direct regulation on limiting content consumption velocity for young users	3	6	8	6	8	6.2
Co-Regulations	3	8	7	5	7	6
Centralised anonymous database	2	9	9	2	8	6
Post-warnings	6	6	5	6	6	5.8
Direct regulations on highly harmful content (e.g. hate speech, incitement to violence)	3	8	6	5	7	5.8
Self-regulation	5	6	5	6	7	5.8
Recommending more often positive content on a topic where other possibly harmful content is visualised in the user feed or home page.	5	5	7	5	7	5.8
Rewarding manual harminformation detection	5	8	7	4	4	5.6

Media literacy campaigns targeting vulnerable communities	4	6	7	6	5	5.6
Reduce the visibility of sensitive content when over confidently shared by non-credible sources	4	6	6	6	5	5.4
Fact checking	7	6	6	3	3	5
Bot detection	6	4	3	5	6	4.8

Table 4 Priorities defined for each proposed or identified solution

A high priority solution does not imply its capability of solving Harminformation, however, it should be an indication of a solution that can be implemented fast, cheap and with a high positive outcome. The positive outcome depends on what problem the solution aims to solve as defined and discussed in chapters 4 and 5.

Solutions have dependencies, in this sense some solutions might share the implementation of software components or comply with the same regulations. For example, the detection of vulnerable communities, virality and credible sources are necessary for the implementation of several solutions. Additionally, whenever citing preventing or monitoring Harminformation there might be the necessity to implement the Harminformation metrics in order to perform the monitoring.

7. Focus

In this section, the focus step of the framework is described, which is crucial for addressing Harminformation solutions in a detailed and prioritised manner. The focus step emphasises the importance of planning, implementing, measuring, and adapting solutions using an agile methodology. This iterative process ensures that interventions remain effective and responsive to evolving challenges. By systematically analysing each solution it is possible to refine and optimise strategies to mitigate harmful content. Importantly, this cycle includes a periodic reassessment to verify the continued relevance and effectiveness of the solutions, ensuring they adapt to changes over time and maintain their efficacy in an ever-changing information landscape.

7.1 Plan-Do-Check-Act

Applying the plan-do-check-act cycle on a real world solution requires a rigorous and detailed planning of the implementation, resources needed, agile methodologies and stakeholders involved. After planning the implementation must be performed and the results measured. Finally, the cycle must be repeated to propose and implement iterative improvements. In this section we briefly simulate the application of the plan-do-check-act cycle on the solutions identified in this thesis. The application of the cycle is short and abstract to avoid overloading the reader with detailed information unnecessary in a simulated environment such as this thesis.

The implementation of strategies to mitigate harmful information involves a structured process, integrating various solutions based on a cycle of planning, implementation, measurement, and reaction. Each solution requires careful consideration of its unique characteristics and intended outcomes. In the list below for each solution an example of the plan-do-check-act cycle is shown. These examples are short and aim to provide an idea of how the cycle could work. However, in a non-simulated environment a much deeper analysis is required for each solution.

Inoculation: the inoculation approach entails planning the creation of educational content that preemptively exposes individuals to diluted forms of Harminformation. This content is then disseminated through various channels, such as social media and educational institutions. The effectiveness of this strategy is measured by assessing the knowledge retention and resilience to Harminformation among the target audience, with subsequent adaptations made based on feedback and emerging trends.

Sharing nudges: introducing friction when sharing content from non-credible sources or overly negative content involves identifying such content and designing mechanisms that introduce friction, like confirmation prompts, to users before they share it. The impact of these interventions is monitored by observing changes in sharing behaviour, and adjustments are made to optimise the balance between user experience and the effectiveness of the friction introduced.

Content velocity reduction: reducing the velocity of content spread, particularly for almost-viral topics. This involves developing algorithms that detect near-viral content and applying measures to slow its spread. The effectiveness of these measures is evaluated by analysing changes in the spread velocity and user engagement metrics.

Preventive and early detection: these mechanisms involve planning the development of algorithms capable of identifying potentially harmful content at an early stage. Once implemented, these algorithms are integrated into existing content moderation systems. Their performance is monitored through metrics such as detection speed and accuracy, and adjustments are made to enhance their efficacy.

Community moderators: community moderators play a crucial role in overseeing online interactions and managing harmful content. Planning involves training these moderators, while implementation assigns them specific responsibilities. The effectiveness of their interventions is measured by the volume of harmful content they flag and remove, with ongoing training provided to ensure they remain effective.

Exposure of vulnerable communities to overly negative content: filters and content warnings can be applied. The success of these measures is assessed by tracking engagement and mental health metrics within these communities, and adjustments are made based on the observed impact.

Social media insurance system: innovative approaches like requiring users to pay an insurance fee, which increases if they disseminate harmful content, involve planning partnerships with insurance companies and monitoring user behaviour. The effectiveness of this model is evaluated by tracking incidents of harmful content dissemination and subsequent adjustments to insurance fees.

Harminformation labelling: labelling of harmful information utilises algorithms to identify and flag various types of harmful content. The implementation involves integrating these systems into content platforms, with their accuracy and user interactions with labelled content being key metrics for evaluation. These systems are continually updated based on performance data.

Detecting polarised groups: detecting polarised groups within social networks involves algorithms that map out group dynamics and content circulation patterns. This information is used to develop interventions aimed at reducing polarisation.

Fact-based journalism: supporting free fact-based journalism involves establishing partnerships with credible journalism outlets and promoting access to factual content. The reach and engagement of this content are monitored to evaluate the initiative's success, with strategies refined as needed.

Regulations: regulatory approaches, both co-regulations and direct regulations, involve collaboration with stakeholders to establish guidelines and standards, which are enforced and monitored for compliance and effectiveness. Self-regulation by platforms and organisations is also encouraged, with the development of best practices and evaluation of adherence to these standards.

Incentive programs: incentivising manual detection of harmful information, such as rewarding individuals who report harmful content, require careful planning and monitoring to ensure effectiveness. Similarly, media literacy campaigns targeting vulnerable communities involve developing and delivering educational content to improve media literacy.

Reducing the visibility of sensitive overconfident shared content: overconfident content when shown by non-credible sources requires algorithms to detect such content and reduce its visibility. The effectiveness of these measures is monitored by examining the reach and engagement metrics, with continuous refinement based on the data collected.

7.2 Reconsider

Addressing Harminformation is a dynamic challenge that evolves with time, as both the tactics used by those spreading harmful content and the nature of the content itself continuously change. A solution effective in mitigating harmful content today may become obsolete as new strategies and technologies emerge.

Therefore, it is crucial to periodically reassess the implemented solutions to ensure their continued effectiveness. This involves setting predefined intervals to measure the outcomes of each solution, analysing whether they still meet the intended objectives, and making necessary adjustments. For example, an algorithm designed to detect a specific type of harmful content may need recalibration or replacement as attackers develop methods to bypass detection.

Similarly, educational campaigns and regulatory measures must adapt to reflect the latest understanding of Harminformation trends and audience behaviours. By systematically reevaluating the effectiveness of these solutions, it becomes possible to discard ineffective approaches, refine existing methods, or even innovate entirely new strategies.

8. Implementing the framework in the EU and the DSA

The effective mitigation of harmful information requires a coordinated effort among multiple stakeholders, including governmental institutions such as the European Commission, independent agencies specializing in information harm (e.g., TrustLab), fact-checking organizations (e.g., Politifact), and major digital platform providers (e.g., Meta, TikTok, X, and Alphabet).

In 2024, the European Union (EU) introduced new regulatory measures under the Digital Services Act (DSA), which established a set of soft-law guidelines for Very Large Online Platforms (VLOPs) such as Facebook, TikTok, Instagram, YouTube, and X. The DSA promotes a model of minimal intervention and voluntary compliance within the digital media sector. However, despite this emphasis on self-regulation, the EU concurrently classifies harmful content as a security threat. This apparent contradiction between voluntary governance and the framing of harmful information as a security risk has been critically examined in the literature (Casero-Ripollés et al., 2023).

Social media platforms, which often derive significant revenue from the virality of user-generated content, including harmful or hateful material, may lack sufficient economic incentives to implement voluntary measures that directly undermine their business models. The discretionary nature of the DSA further complicates this issue, as it does not impose explicit mandates on digital platforms but rather leaves content governance decisions to individual companies. This regulatory ambiguity introduces the risk of non-compliance and enables platforms to evade responsibility for the dissemination of harmful information.

The conceptual framework proposed in this thesis could be adopted by the EU to establish a procedural mechanism for addressing Harminformation. The proposed framework aims to provide clearer regulatory expectations for implementing entities, particularly VLOPs, while maintaining alignment with the voluntary principles of the DSA. By integrating both general harm-related issues and platform-specific challenges, the framework would enhance the effectiveness of regulatory interventions.

The implementation of the proposed framework within the EU could follow a structured four-phase cycle: Identify and Define, Prioritize, Focus, and Reconsider (IDPF).

Identify and Define: Every two years, the EU would request each member state to independently and anonymously identify pressing issues related to harmful information. The identification process should involve collaboration between the Ministry of Defense, academia, non-profit organizations, and private sector stakeholders (e.g., defense contractors) to compile a comprehensive list of emerging concerns. Following the collection and dissemination of identified issues, all relevant stakeholders would be invited to propose potential solutions, ideally based on empirical evidence. These proposed solutions would then be submitted to the respective national Ministry of Defense, which would assess their relevance and determine whether to forward them to the EU or discard them. Additionally, member states would have the option to recommend modifications or eliminations of proposed solutions.

Prioritize: Once national Ministries of Defense receive the compiled solutions, they would forward them to the EU, which would subsequently distribute them to all member states for further evaluation. Within a predefined timeframe (e.g., two months), each national Ministry of Defense would assign priority scores to the proposed solutions, following a structured decision-making process (as discussed

in section 6). This prioritization should be informed by continuous engagement with academia, non-profit organizations, and industry representatives. The EU would then aggregate the priority scores, calculating an average ranking to determine the most urgent solutions. Additionally, each member state would be required to provide estimated implementation timelines for each solution, from which the EU would derive a median timeframe. In cases where solutions are recommended for elimination or modification, a majority consensus among member states would be required before formalizing the decision.

Focus: In this phase, the EU would select ten of the highest-priority solutions for implementation and formally communicate the implementation requirements to the designated responsible entities (most often VLOPs but potentially other actors such as public institutions or non-profits). The selected solutions would become legally binding, and implementing entities would be obligated to adhere to the prescribed timelines. In accordance with the Plan-Do-Check-Act (PDCA) cycle and Agile methodologies, implementing entities would be required to periodically review the effectiveness of adopted solutions, iteratively improve them, and submit assessment reports to the EU and national Ministries of Defense. Additionally, any modifications or eliminations of implemented solutions must be addressed as part of the continuous improvement process.

Reconsider: Given the cyclical nature of the IDPF process, each subsequent cycle would provide an opportunity for member states to reassess previously identified issues, evaluate the effectiveness of implemented solutions, and propose necessary modifications or eliminations. The evaluation should be based on analysis coming from the implementation party and peer reviewed scientific articles assessing the effectiveness of a solution. This iterative approach ensures adaptability to the evolving nature of harmful information and allows for continuous refinement of regulatory interventions.

This implementation of the framework offers a structured approach for addressing the challenges associated with Harminformation within the EU while maintaining the voluntary principles of the DSA. The timeframes outlined for each phase of the IDPF cycle serve as an illustrative example, subject to adjustment based on empirical findings and policy considerations. While VLOPs will likely serve as the primary implementation entities, certain solutions, such as media literacy initiatives, may necessitate leadership from public institutions or non-profits rather than private corporations.

Finally, similar to the DSA's enforcement mechanisms, non-compliance with mandated solutions should result in financial penalties for private companies, ensuring accountability among implementing entities.

9. Conclusion

In this thesis a complete methodology to counteract Harminformation has been proposed. A high level conceptual framework to define a process stakeholders should follow to counteract information harm has been designed; detailed information on the different types and characteristics of information harm, solutions to counteract them, metrics to define them formally and detect Harminformation have been proposed. The framework and the detailed problems and solutions are vast and will require constant effort to be implemented. Furthermore, the framework and solutions have limitations that need to be addressed. In this section, guidelines on the implementation and limitations are described together with final remarks of this thesis.

9.1 Limitations and future work

This thesis addresses the complex issue of defining Harminformation and introducing a multiparty process to identify, describe and implement the most effective solutions. It does so by dissecting the confusion on the meanings of terms used in existing literature and proposing solutions aimed at counteracting information harm.

The diverse nature of harmful content necessitates real-life experimentation with data that is predominantly owned by large corporations like Meta and TikTok. A significant limitation of this study is that, although the proposed solutions are inspired by well-established research, certain approaches, such as the single content and aggregated topic metrics, have yet to be empirically tested.

The proposed framework adopts an agile process approach that could be considered by institutions like the European Commission in their efforts to combat Harminformation. The framework's generality is intentional, as it avoids delving into the specifics of each solution. Instead, it emphasises the procedural aspects that stakeholders should follow, given the constantly evolving nature of Harminformation. This approach allows for adaptability and responsiveness to new challenges. The prioritisation outlined in this thesis is based on the judgement of a single individual, incorporating both personal intuition and supporting research. The aim is to show a potential method for a group of experts to collaboratively decide on priorities and reach consensus, despite the inherently subjective nature of such assessments.

Although a range of solutions is discussed, this thesis does not claim to address all types of harmful information. Instead, it focuses on contributing to the most contentious issues, such as balancing free speech with the regulation of harmful content, promoting deep thinking, discouraging the creation and dissemination of sensationalist and emotionally charged content, and the early detection and prevention of harmful material. This selective focus is intended to offer practical solutions to these particularly challenging areas, rather than attempting an exhaustive treatment of the entire scope of Harminformation.

One of the main limitations of this work is the absence of validation and evaluation of the proposed conceptual framework under real-world conditions, including testing within national and international institutions (e.g., the EU). Future works should focus on engaging one or more public institutions with legislative authority to implement the framework, enabling an assessment of its strengths and weaknesses in practical contexts. The effectiveness of individual Harminformation solutions approved through the framework should be measured using tailored quantitative analyses, validated through

peer-reviewed studies. Additionally, after several years of experimental adoption by a public institution, the overall impact of the framework could be evaluated by comparing the effectiveness of proposed and implemented measures against different types of Harminformation to a no-framework baseline scenario where existing legislation and practices are applied.

9.2 Final remarks

In the contemporary information landscape, the rapid dissemination of information, particularly on social media platforms, has given rise to the concerning phenomenon of disinformation. The proliferation of disinformation poses threats to various facets of society, including the erosion of trust in information sources, the manipulation of public opinion, and the potential to incite social and political unrest. The term disinformation has been cause of confusion in the research community because misinformation, disinformation, and malinformation meanings easily overlap as users in social media unintentionally share false information causing harm to the public's opinion. Therefore, in this work the term Harminformation is defined to solve this issue. Harminformation can be easily defined as harmful information. In this work this new term has been defined abstractly to conceptualise the phenomena that is necessary to counteract. However, to avoid confusion and errors caused by natural language interpretations the most important types of Harminformation have been listed and described adding a layer of specialisation to the abstract definition.

This work contribution to research is manifold: 1) The term Harminformation is defined, characterised in Harminformation types and compared with the commonly accepted definition of Disinformation; 2) An agile conceptual framework indicating a process stakeholders could follow to implement solutions has been proposed; 3) Harminformation related problems described in the literature have been identified; 4) Solutions to these problems have been proposed (both existing and new solutions); 5) Precise metrics to detect Harminformation have been described and the way to calculate them has been written in pseudocode, together with citations of works proposing or using AI algorithms needed to compute them; 6) The process of solutions prioritisation among stakeholders is simulated; 7) Finally, the step of implementation, measurement and reactions is described for each solution.

This thesis therefore aims to provide a multi-stakeholder and multidisciplinary guide to fight information harm. How and who spreads Harminformation, why it is harmful, who is more vulnerable to harmful content and why we are unable to spot it are the topics faced and described in detail. These topics and problems are based on psychological, cognitive, law and technological works and results in the literature. To each of these problems solutions are detailed with their pros and cons. Metrics to detect Harminformation, both at a single content level and topic level (i.e. as an aggregation of single contents representing a topic), are proposed; finally, adaptations to the metrics, limitations of the work and future directions are outlined. These include the necessity of constant cooperation among the stakeholders and the implementation of many solutions at different societal levels only after the comprehension of such a complex phenomena.

In summary, the contribution of this work is represented by one of the first attempts in the research community to provide a holistic and clear view of the complex phenomena of information harm starting from the proposal of a framework for stakeholders, to the precise definition of Harminformation, Harminformation types and the ways to counteract them.

References

Aichner, T., Grünfelder, M., Maurer, O., & Jegeni, D. (2021). Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking, 24*(4), 215-222. DOI: 10.1089/cyber.2020.0134

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, *13*(1), 30.

Aisch G, Huang J, and Kang C (2016). Dissecting the #PizzaGate conspiracy theories. Nytimes.com. Available at: https://www.nytimes.com/interactive/2016/12/10/business/media/pizzagate.html.

Akinpelu, S., Viriri, S., & Adegun, A. (2023). Lightweight deep learning framework for speech emotion recognition. *IEEE Access*.

Albarracin D, Romer D, Jones C, et al. (2018). Misleading claims about tobacco products in YouTube videos: experimental effects of misinformation on unhealthy attitudes. Journal of Medical Internet Research, 20(6), e229.

Allcott H, and Gentzkow M (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 31(2), 211–236.

Altay, S., Nielsen, R. K., & Fletcher, R. (2024). News can help! The impact of news media and digital platforms on awareness of and belief in misinformation. *The International Journal of Press/Politics*, *29*(2), 459-484.

Alter, A. L. (2017). *Irresistible: The rise of addictive technology and the business of keeping us hooked*. Penguin Press.

Knight Foundation (2020). Trust, media and democracy. https://knightfoundation.org/wp-content/uploads/2020/08/American-Views-2020-Trust-Media-and-Democracy.pdf

Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., ... & Rand, D. G. (2023). Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, *7*(9), 1502-1513.

Avin, C., Daltrophe, H., & Lotker, Z. (2024). On the impossibility of breaking the echo chamber effect in social media using regulation. Scientific Reports, 14(1), 1107.

Taylor, J. Butler, J. (2024). Australian prime minister labels Elon Musk ‘an arrogant billionaire who thinks he is above the law.’ The Guardian. https://amp.theguardian.com/australia-news/2024/apr/23/elon-musk-anthony-albanese-x-twitter-australia-prime-minister-sydney-church-stabbing-videos-ntwnfb

Bad News Game (2017). Available at: https://getbadnews.com/#intro (accessed 5 April 2019).

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature reviews neuroscience*, *16*(7), 419-429.

Bastos, M. T., & Mercea, D. (2018). The public accountability of social platforms: Lessons from a study on bots and trolls in the Brexit campaign. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20180003. DOI: 10.1098/rsta.2018.0003

Bak-Coleman, J., et al. (2021). Combating misinformation requires a multi-disciplinary approach. *Science, 373*(6562), 658-660. https://doi.org/10.1126/science.abi6864

Bennett, W. L., & Livingston, S. (2020). *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States*. Cambridge University Press.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Berger, J., & Milkman, K. L. (2012). What makes online content viral?. *Journal of marketing research*, *49*(2), 192-205.

Bernal, P. (2021). Social media regulation: Why we must ensure it is done right. *Communications Law*, 26(1), 1-15.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bollen J, Mao H, and Zeng X (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

Bond Jr, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and social psychology Review*, *10*(3), 214-234.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information sciences*, *497*, 38-55.

Bontcheva, K., & Posetti, J. (2020). Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. *UNESCO*.

Bottoms, H. C., Eslick, A. N., & Marsh, E. J. (2010). Memory and the Moses illusion: Failures to detect contradictions with stored knowledge yield negative memorial consequences. *Memory*, *18*(6), 670-678.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. Nature communications, 10(1), 7.

Brady WJ, Wills JA, Jost JT, Tucker JA, and Van Bavel JJ (2017). Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28), 7313–7318.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978-1010.

Brashier, N. M., Umanath, S., Cabeza, R., & Marsh, E. J. (2017). Competing cues: Older adults rely on knowledge in the face of fluency. *Psychology and aging*, *32*(4), 331.

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. Annals of the International Communication Association, 48(2), 139–166. https://doi.org/10.1080/23808985.2024.2323736

Bursztyn L, Petrova M, Enikolopov R, et al. (2018). Social media and xenophobia: evidence from Russia. The American Economic Association's RCT Registry. Available at: https://doi.org/10.1257/rct.3066-1.0.

Buczel, K. A., Siwiak, A., Szpitalak, M., & Polczyk, R. (2024). How do forewarnings and post-warnings affect misinformation reliance? The impact of warnings on the continued influence effect and belief regression. Memory & Cognition, 1-17.

Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content-based fake news detection with machine and deep learning: a systematic review. *Neurocomputing*, *530*, 91-103.

Capuano, N., Meyer, M. & Nota, F.D. (2024). Analyzing the impact of conversation structure on predicting persuasive comments online. *J Ambient Intell Human Comput*.

Card, D., Boydstun, A., Gross, J. H., Resnik, P., & Smith, N. A. (2015, July). The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 438-444).

Carr, N. (2010). The Shallows: What the Internet Is Doing to Our Brains. W. W. Norton & Company. Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Casero-Ripollés, A., Tuñón, J., & Bouza-García, L. (2023). The European approach to online disinformation: geopolitical and regulatory dissonance. *Humanities and Social Sciences Communications*, *10*(1), 1-10.

Cavaliere, D., Fenza, G., Loia, V., & Nota, F. (2023). Emotion-Aware Monitoring of Users' Reaction With a Multi-Perspective Analysis of Long-and Short-Term Topics on Twitter.

Chadwick, A., & Stanyer, J. (2022). Deception as a bridging concept in the study of disinformation, misinformation, and misperceptions: Toward a holistic framework. Communication Theory, 32(1), 1-24.

Chee F (2020). Combat 5G COVID-19 fake news, urges Europe. US. Available at: https://www.reuters.com/article/us-eu-telecoms-5g/combat-5g-covid-19-fake-news-urges-europe-idUSKBN2392N8.

Chen E, Kao G, and Deibert R (2020). Beyond infiltration: Understanding the scale and scope of Chinese government influence operations on Twitter. Media International Australia, 178(1), 38–54.
Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014, April). Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web* (pp. 925-936).

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.

Cock Buning, M. (2018). 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.

Code of Practice Report on Disinformation | TrustLab. (n.d.). Www.trustlab.com. Retrieved June 30, 2024, from https://www.trustlab.com/resources/codeofpractice-disinformation

Comito, C., Forestiero, A., & Pizzuti, C. (2019, October). Word embedding based clustering to detect topics in social media. In *IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 192-199).

Coombs WT (2014). Ongoing Crisis Communication: Planning, Managing, and Responding. Sage Publications.

Cook, J., Ecker, U. K., Trecek-King, M., Schade, G., Jeffers-Tracy, K., Fessmann, J., ... & McDowell, J. (2023). The cranky uncle game—combining humor and gamification to build student resilience against climate misinformation. *Environmental Education Research*, *29*(4), 607-623.

Dame Adjin-Tettey, T. (2022). Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. *Cogent arts & humanities*, *9*(1), 2037229.

Daniel K (2017). Thinking, fast and slow.

Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017, January). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)* (pp. 86-95).

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Department for Science, Innovation & Technology. (2024, May 8). Online Safety Act: explainer. GOV.UK. https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dikovitsky, V. V., & Fedorov, A. M. (2020). Topic Clustering of Social Media Using Multilayer Text Analysis. In *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4* (pp. 931-938). Springer International Publishing.

DiResta R, Shaffer K, Ruppel M, Sullivan B, Matney R, Fox A, and Roberts C (2020). The Tactics & Tropes of the Internet Research Agency. New Knowledge.

Dreyfuss, E. (2017). Want to make a lie seem true? Say it again. And again. And again. Wired. com, 11.

DSA Transparency Database. (n.d.). Transparency.dsa.ec.europa.eu. Retrieved July 1, 2024, from https://transparency.dsa.ec.europa.eu/

Durach, F., Bârgăoanu, A., & Nastasiu, C. (2020). Tackling disinformation: EU regulation of the digital space. *Romanian journal of European affairs*, *20*(1).

Dutton, W. H., Reisdorf, B., Dubois, E., & Blank, G. (2017). Social shaping of the politics of internet search and networking: Moving beyond filter bubbles, echo chambers, and fake news.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29.

Ecker, U., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation poses a bigger threat to democracy than you might think. *Nature*, *630*(8015), 29-32.

Elmas, T., Stephane, S., & Houssiaux, C. (2023, April). Measuring and detecting virality on social media: the case of twitter's viral tweets topic. In *Companion Proceedings of the ACM Web Conference 2023* (pp. 314-317).

Emerson, A., Houghton, P., Chen, K., Basheerabad, V., Ubale, R., & Leong, C. W. (2022, November). Predicting User Confidence in Video Recordings with Spatio-Temporal Multimodal Analytics. In *Companion Publication of the 2022 International Conference on Multimodal Interaction* (pp. 98-104).

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540-551.

European Commission (2018a). Fake news and online disinformation. Digital Single MarketPolicy. Available at:
https://ec.europa.eu/digital-single-market/en/fake-news-disinformation.

European Commission (2018b). Tackling Online Disinformation: A European Approach. Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/report-implementation-communication-tackling-online-disinformation-european-approach](https://ec.europa.eu/digital-single-market/en/news/report-implementation-communication-tackling-online-disinformation-european-approach.

European Commission. (2022). Digital Services Act. Available at: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

Ferrara, E. (2017). Disinformation and Social Bot Operations in the Run up to the 2017 French Presidential Election. *First Monday*, 22(8).

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96-104.

Flamino, J., Szymanski, B. K., Bahulkar, A., Chan, K., & Lizardo, O. (2021). Creation, evolution, and dissolution of social groups. Scientific reports, 11(1), 17470.

Flood, M. (2009). The harms of pornography exposure among children and young people. *Child Abuse Review: Journal of the British Association for the Study and Prevention of Child Abuse and Neglect*, *18*(6), 384-400.

Fogg, B. J. (2003). *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann.

Fridman, J., Barrett, L. F., Wormwood, J. B., & Quigley, K. S. (2019). Applying the theory of constructed emotion to police decision making. *Frontiers in psychology*, *10*, 1946.

Friggeri A, Adamic LA, Eckles D, and Cheng J (2014). Rumor Cascades. Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM).

Ghanem, B., Rosso, P., & Rangel, F. (2020). An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)*, *20*(2), 1-18.

Ghosal, T., Saikh, T., Biswas, T., Ekbal, A., & Bhattacharyya, P. (2022). Novelty detection: A perspective from natural language processing. *Computational Linguistics*, *48*(1), 77-117.

Gruhl, C., Sick, B., & Tomforde, S. (2021). Novelty detection in continuously changing environments. *Future Generation Computer Systems*, *114*, 138-154.

González Esparza, H. J., Florencia, R., Díaz Román, J. D., & Mendoza-Carreón, A. (2023). Text Classifier of Sensationalist Headlines in Spanish Using BERT-Based Models. In *Innovations in Machine and Deep Learning: Case Studies and Applications* (pp. 109-131). Cham: Springer Nature Switzerland.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1). https://doi.org/10.1177/2053951719897945

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (pp. 369-376).

Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, *10*.

Guess AM, Nyhan B, and Reifler J (2019). Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign. European Research Council.

Guess, A. M., Nagler, J., & Tucker, J. (2019). Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook. *Science Advances*, 5(1), eaau4586.

Guo, L., & Vargo, C. (2020). "Fake news" and emerging online media ecosystem: An integrated intermedia agenda-setting analysis of the 2016 US presidential election. *Communication Research*, *47*(2), 178-200.

Hao K, Chen Y, and Zuo X (2018). The Content and Focus of State-Sponsored Propaganda in China: Evidence from Twitter. SSRN Electronic Journal.
Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. Cognitive Research: Principles and Implications, 6, Article 38.

Hao, K. (2018). Even the best AI for spotting fake news is still terrible. *MIT Technology Review, 03-out-2018*.

Hayawi, K., Saha, S., Masud, M. M., Mathew, S. S., & Kaosar, M. (2023). Social media bot detection with deep learning methods: a systematic review. *Neural Computing and Applications*, *35*(12), 8903-8918.

Heiskala, R., & Aro, J. (Eds.). (2018). *Policy design in the European Union: an empire of shopkeepers in the making?*. Springer.

Helfmann, L., Djurdjevac Conrad, N., Lorenz-Spreen, P., & Schütte, C. (2023). Modelling opinion dynamics under the impact of influencer and media strategies. *Scientific Reports*, *13*(1), 19375.

Highsmith, J. (2009). Agile project management: creating innovative products. Pearson education.

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. Nature Human Behaviour, 1-9.

Hotten R (2015). Volkswagen: the scandal explained. Available at: https://www.bbc.com/news/business-34324772.

Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to online disinformation: A framework for cross-national comparative research. The international journal of press/politics, 25(3), 493-516.

Hung, T. C., & Hung, T. W. (2022). How China's cognitive warfare works: A frontline perspective of Taiwan's anti-disinformation wars. Journal of Global Security Studies, 7(4), ogac016.

InVID Project (2017). Web Lyzard technology. Available at: https://invid.weblyzard.com/.

Jang, S. M., & Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in human behavior*, *80*, 295-302.

Jeong, S. H., Cho, H., & Hwang, Y. (2012). Media literacy interventions: A meta-analytic review. *Journal of communication*, *62*(3), 454-472.

Jolley D, and Douglas K (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. PLoS ONE, 9(2), e89177.

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American behavioral scientist*, *65*(2), 371-388.

Jurkowitz, M., Mitchell, A., Shearer, E., & Walker, M. (2020). US media polarization and the 2020 election: A nation divided. *Pew Research Center*, *24*.

Kazeem Y (2018). Nigerian media houses are forming a coalition to combat fake news ahead of next year's elections. Quartz Africa. Available at: https://qz.com/africa/1478737/fake-news-media-collaborate-ahead-of-nigeria-2019-election/.

Khan, M. S., Malik, M. S. I., & Nadeem, A. (2024). Detection of violence incitation expressions in Urdu tweets using convolutional neural network. *Expert Systems with Applications*, *245*, 123174.

Khan, T., Sarkar, R., & Mollah, A. F. (2021). Deep learning approaches to scene text detection: a comprehensive review. *Artificial Intelligence Review*, *54*, 3239-3298.

Keles, B., McCrae, N., & Grealish, A. (2020). A systematic review: The influence of social media on depression, anxiety, and psychological distress in adolescents. *International Journal of Adolescence and Youth, 25*(1), 79-93. DOI: 10.1080/02673843.2019.1590851

Kemeny, J. G. (1952). WV Quine. Two dogmas of empiricism. The philosophical review, vol. 60 (1951), pp. 20–43. *The Journal of Symbolic Logic*, *17*(4), 281-283.

King's College and Ipsos MORI (2018). Brexit misperceptions. Report, London. Available at: https://ukandeu.ac.uk/wp-content/uploads/2018/10/Brexit-misperceptions.pdf.

Klingberg T (2009). The overflowing brain: Information overload and the limits of working memory. Oxford University Press.

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, *120*(7), e2210666120.

Kumar, H., & Martin, A. (2023). Artificial Emotional Intelligence: Conventional and deep learning approach. *Expert Systems with Applications*, *212*, 118651.

Kušen E, and Strembeck M (2018). Politics, sentiments, and misinformation: an analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. Online Social Networks and Media—Journal, 5, 37–50.

Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long‐term memory. *Cognitive science*, *10*(4), 477-493.

Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, and Zittrain JL (2018). The science of fake news. Science, 359(6380), 1094–1096.

Lazer, D., Baum, M., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017). Combating fake news: An agenda for research and action.

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, *13*.

Leonard, T. C. (2008). Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness: Yale University Press, New Haven, CT, 2008, 293 pp, $26.00.

Lewandowsky S, Ecker UKH, and Cook J (2012). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. Journal of Applied Research in Memory and Cognition, 6(4), 353–369.

Lewandowsky, S., Cook, J., Ecker, U., Albarracin, D., Kendeou, P., Newman, E. J., ... & Zaragoza, M. S. (2020). The debunking handbook 2020.

Loomba, S., De Figueiredo, A., Piatek, S. J., De Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour*, *5*(3), 337-348.

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, *118*(23), e2019527118.

Lustig, C., & Meck, W. (2009). The overflowing brain: Information overload and the limits of working memory. New England Journal of Medicine, 360(14), 1469.

Mach, K. J., Salas Reyes, R., Pentz, B., Taylor, J., Costa, C. A., Cruz, S. G., ... & Klenk, N. (2021). News media coverage of COVID-19 public health and policy information. *Humanities and Social Sciences Communications*, *8*(1).

Maréchal, N., & Biddle, E. R. (2020). It's not just the content, it's the business model: Democracy's online speech challenge. *Washington, DC: Ranking Digital Rights, New America Foundation, March*.

Marcial-Basilio, J. A., Aguilar-Torres, G., Sánchez-Pérez, G., Toscano-Medina, L. K., & Perez-Meana, H. M. (2011). Detection of pornographic digital images. *International journal of computers*, *5*(2), 298-305.

Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 101710.

Marwick A, and Lewis R (2017). Media Manipulation and Disinformation Online. Data Society Research Institute.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, *114*(48), 12714-12719.

McLoughlin, K. L., Brady, W. J., & Crockett, M. J. (2021). The role of moral outrage in the spread of misinformation. *TMS Proceedings*.

Meel, P., & Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, *153*, 112986.

Meta. (2023). Continuing our Approach to Political Content on Instagram and Threads. Available at: https://about.instagram.com/blog/announcements/continuing-our-approach-to-political-content-on-instagram-and-threads/

Mirhoseini, M., Early, S., El Shamy, N., & Hassanein, K. (2023). Actively open-minded thinking is key to combating fake news: A multimethod study. *Information & Management*, *60*(3), 103761.

Modha, S., et al. (2021). Overview of the HASOC subtrack at fire 2021: hate speech and offensive content identification in English and Indo-Aryan languages and conversational hate speech. In *Forum for Information Retrieval Evaluation*, pp. 1–3.

Moen, R., & Norman, C. (2006). Evolution of the PDCA cycle.

Mohammadi, S., Perina, A., Kiani, H., & Murino, V. (2016). Angry crowds: Detecting violent events in videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14* (pp. 3-18). Springer International Publishing.

Mondal, S., Pal, S., Saha, S. K., & Chanda, B. (2017, December). Violent/non-violent video classification based on deep neural network. In *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)* (pp. 1-6). IEEE.

Montag, C., & Diefenbach, S. (2020). Towards Homo Digitalis: Important Research Issues for Psychology and the Neurosciences at the Dawn of the Internet of Things and the Digital Society. *Sustainability, 12*(18), 7412. DOI: 10.3390/su12187412

Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, *57*(3), 102087.

Müller K, and Schwarz C (2017). Fanning the flames of hate: social media and hate crime. SSRN Electronic Journal. Available at: http://dx.doi.org/10.2139/ssrn.3082972.

Naeem, B., Khan, A., Beg, M. O., & Mujtaba, H. (2020). A deep learning framework for clickbait detection on social area network using natural language cues. *Journal of Computational Social Science*, *3*(1), 231-243.

Naik, A. J., & Gopalakrishna, M. T. (2021). Deep-violence: individual person violent activity detection in video. *Multimedia Tools and Applications*, *80*(12), 18365-18380.

Nemeth, C., & Bonneau, C. (2018). The Effects of Regulating Social Media Content on Mental Health: An Empirical Analysis. *Journal of Social Media in Society, 7*(2), 234-251.

Nguyen, V. H., Sugiyama, K., Nakov, P., & Kan, M. Y. (2020, October). Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 1165-1174).

Nova, D., Ferreira, A., & Cortez, P. (2019). A machine learning approach to detect violent behaviour from video. In *Intelligent Technologies for Interactive Entertainment: 10th EAI International Conference, INTETAIN 2018, Guimarães, Portugal, November 21-23, 2018, Proceedings 10* (pp. 85-94). Springer International Publishing.

Nygren, T., & Guath, M. (2022). Students evaluating and corroborating digital news. *Scandinavian Journal of Educational Research*, *66*(4), 549-565.

Obar, J. A., & Wildman, S. S. (2015). Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy, 39*(9), 745-750. DOI: 10.1016/j.telpol.2015.07.014

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*(3), 999-1015.

Pariser E (2011). The filter bubble: How the new personalized web is changing what we read and how we think. Penguin.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. Journal of experimental psychology: general, 147(12), 1865.

Pennycook G, and Rand DG (2018). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings. Management Science.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388-402.

Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., ... & Rocha, A. (2017). Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, *230*, 279-293.

Politifact (2007). Available at: https://www.politifact.com/.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Potter, W. J. (2018). *Media literacy*. Sage publications.

Prebunking with Google. (n.d.). https://prebunking.withgoogle.com/

Prior, M., Sood, G., & Khanna, K. (2015). You Cannot be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions. Quarterly Journal of Political Science, 10(4), 489–518.

Que, X., Checconi, F., Petrini, F., & Gunnels, J. A. (2015, May). Scalable community detection with the louvain algorithm. In *2015 IEEE international parallel and distributed processing symposium* (pp. 28-37). IEEE.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118.

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & Van Der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis) information. *Nature Human Behaviour*, *7*(6), 892-903.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., ... & Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Renda A (2018). The legal framework to address "fake news": Possible policy actions at the EU level. European Parliament.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*(1), 1-10. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G.,. & Van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science*, *7*(10), 201199.

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.

Rosling, H. (2023). *Factfulness*. Flammarion.

Ross, L., & Ward, A. (1996). Naive Realism in Everyday Life: Implications for Social Conflict and Misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), Values and Knowledge (pp. 103–135). Mahwah: Lawrence Erlbaum.

Saltor, J., Barberia, I., & Rodríguez‐Ferreiro, J. (2023). Thinking disposition, thinking style, and susceptibility to causal illusion predict fake news discriminability. *Applied Cognitive Psychology*, *37*(2), 360-368.

Saha, S., Junaed, J. A., Saleki, M., Rahouti, M., Mohammed, N., & Amin, M. R. (2023, December). Blp-2023 task 1: Violence inciting text detection (vitd). In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)* (pp. 365-375).

Serhan Y (2018). Italy scrambles to fight misinformation ahead of its elections. The Atlantic. Available at: https://www.theatlantic.com/international/archive/2018/02/europe-fake-news/551972/.

Sgambati, T. J., & Ayduk, O. N. (2023). Intellectual Humility and Political Polarization: An Exploration of Social Networks, Attitudes, and Affect. *Political Psychology*, *44*(4), 807-828.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22-36.

Silverman C (2016). This analysis shows how viral fake election news stories outperformed real news on Facebook. Buzzfeed News. Available at: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

Smith, J., Legg, P., Matovic, M., & Kinsey, K. (2018). Predicting user confidence during visual decision making. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(2), 1-30.

Solovev, K., & Pröllochs, N. (2022, April). Moral emotions shape the virality of COVID-19 misinformation on social media. In *Proceedings of the ACM web conference 2022* (pp. 3706-3717).

Snopes (1994). Available at: http://www.snopes.com/.

Starbird K, Arif A, Wilson T, Stanek S, and Ashley S (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 104.

Starbird K, Maddock J, Orand M, et al. (2014). Rumors, false flags, and digital vigilantes: misinformation on twitter after the 2013 Boston marathon bombing. In: iConference 2014 proceedings, Berlin, Germany, 4–7 March, pp. 654–662. DOI: 10.9776/143

Stone, J. V. (2015). Information theory: a tutorial introduction.

Tanvir, A. A., Mahir, E. M., Huda, S. A., & Barua, S. (2020, January). A hybrid approach for identifying authentic news using deep learning methods on popular Twitter threads. In *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)* (pp. 1-6). IEEE.

Tarski, A. (2021). The semantic conception of truth and the foundations of semantics.

Tsfati, Y., & Cappella, J. N. (2003). Do people watch what they do not trust? Exploring the association between news media skepticism and exposure. Communication Research, 30(5), 504-529.

Twenge, J. M., & Campbell, W. K. (2019). Media use is linked to lower psychological well-being: Evidence from three datasets. *Psychiatric Quarterly, 90*(2), 311-331. DOI: 10.1007/s11126-019-09630-7

Vaidhyanathan, S. (2018). *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford University Press.

Vahedi, Z., Sibalis, A., & Sutherland, J. E. (2018). Are media literacy interventions effective at changing attitudes and intentions towards risky health behaviors in adolescents? A meta-analytic review. *Journal of adolescence*, *67*, 140-152.
Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*, 395-416.

van Prooijen, J. W., & van Vugt, M. (2018). Conspiracy Theories: Evolved Functions and Psychological Mechanisms. *Perspectives on Psychological Science*, 13(6), 770-788.

Vannucci, A., & Ohannessian, C. M. (2019). Social media use subgroups differentially predict psychosocial well-being during early adolescence. *Journal of Youth and Adolescence, 48*(8), 1469-1493. DOI: 10.1007/s10964-019-01051-1

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Wagner, B., Rozgonyi, K., Sekwenz, M. T., Cobbe, J., & Singh, J. (2020, January). Regulating transparency? Facebook, twitter and the German network enforcement act. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 261-271).

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

Weng, L., Menczer, F., & Ahn, Y. Y. (2013). Virality prediction and community structure in social networks. *Scientific reports*, *3*(1), 1-6.

Williams, E., Gray, J., Morris, E., Bradshaw, B., Williams, K., & Dixon, B. (2016, October). A comparison of two methods for the topical clustering of social media posts. In *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 1-7). IEEE.

Wijesekara, M., & Ganegoda, G. U. (2020, September). Source credibility analysis on Twitter users. In *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)* (pp. 96-102). IEEE.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in Social Media: Definition, Manipulation, and Detection. *ACM SIGKDD Explorations Newsletter*, 21(2), 80-90.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in neural information processing systems*, *32*.

Zhang, X. S., Wang, R. S., Wang, Y., Wang, J., Qiu, Y., Wang, L., & Chen, L. (2009). Modularity optimization in community detection of complex networks. *Europhysics Letters*, *87*(3), 38002.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1-40.

Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech detection. *IEEE Access*, *8*, 128923-128929.