Human-Centric Cybersecurity Partnership

2025
SUMMER

# MISINFORMATION & CYBERSECURITY

## EXAMINING THE THREAT OF MISINFORMATION TO DIGITAL MEDIA

SARAH ABDI
MUHAMMAD ZAID ARIF
TEHYA BLAKE
ARIELLE CHAÎNÉ
GETHO OXEUS
CASSIDY WEEKES

## HUMAN-CENTRIC CYBERSECURITY REPORT PROJECT

The 2024 Human-Centric Cybersecurity Report Project brought together postgraduate students from across Canada and the United Kingdom to work with our partners from both private industry and the public sector to produce reports looking at the problem of ransomware through a transdisciplinary lens.

## ABOUT HC2P

The Human-Centric Cybersecurity Partnership (HC2P) is a transdisciplinary group of scholars, government, industry and not-for-profit partners that generate research and mobilize knowledge that will help create a safer, more secure, more democratic and more inclusive digital society.

## ACKNOWLEDGEMENTS

We would like to thank the Canadian Centre for Cyber Security (CCCS), the Canadian Cyber Threat Exchange (CCTX), the Canadian Security Intelligence Service (CSIS), Cybeeco, Desjardins, Flare Systems, Google Cloud, Innovation, Science and Economic Development Canada (ISED), the Institute for Data Valorization (IVADO), the National Bank of Canada, the National Cybercrime Coordination Centre (NC3), the National Research Council of Canada (NRC), Public Safety Canada, the Royal Canadian Mounted Police (RCMP), Statistics Canada, the University of Montreal and the University of Ottawa for their efforts in supporting this project.

We also acknowledge the valuable contributions of academic collaborators Alina Dulipovici, Andreea Musulan, Anne Broadbent, Florian Martin-Bariteau, Frederic Schlackl, J. Marshall Palmer and Philippe Lamontagne.



Cite as:

Abdi, S., Arif, M. Z., Blake, T., Chaîné, A., Oxeus, G., & Weekes, C. (2025). Misinformation & cybersecurity: Examining the threat of misinformation to digital media. Human-Centric Cybersecurity Partnership (HC2P).

Dépôt légal,

ISBN: 978-1-7387249-8-7

The Human-Centric Cybersecurity Partnership is supported in part by funding from the Social Sciences and Humanities Research Council.

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

# Table of Contents

# Misinformation and Cybersecurity

EXAMINING THE THREAT OF MISINFORMATION TO DIGITAL MEDIA

## Executive Summary

This report explores the critical issue of misinformation and its impact on cybersecurity, focusing on political, regulatory, societal, and technological dimensions. The rapid proliferation of misinformation, particularly on social media, poses significant challenges to individuals, organisations, and governments alike. The report provides a detailed examination of the evolution of misinformation, the factors contributing to its spread, and the vulnerabilities it creates within cybersecurity frameworks.

The increasing use of social media as a news source has facilitated the rapid dissemination of false information, exacerbated by cognitive biases and social reinforcement mechanisms. Political misinformation has influenced public opinion, heightened polarization, and undermined democratic processes. Furthermore, misinformation surrounding health issues, such as the COVID-19 pandemic, has had dire consequences, including reduced vaccine uptake and the spread of dangerous health practices.

Technological advancements, including Artificial Intelligence (AI) and deepfakes, have further complicated efforts to detect and mitigate

misinformation. The sophistication of AI-generated content makes it challenging for detection tools to keep pace, raising concerns about the authenticity of digital information.

Countermeasures discussed in the report include proactive strategies like prebunking and educational initiatives, as well as reactive approaches such as fact-checking, automated content labeling, and AI-driven detection. While these countermeasures have shown promise, they also face limitations. Prebunking, for instance, may be less effective in the varied and fast-paced environments of social media, and fact-checking often struggles to counteract the emotional and social factors that drive misinformation sharing.

The report concludes by recommending a multifaceted approach to combat misinformation. This includes regulatory reforms that hold platforms accountable for the content they promote, enhanced public digital literacy programs, and the continued development of advanced AI tools to detect false content. Collaboration between governments, industry, and civil society is essential to create a more resilient digital information ecosystem that protects against the harms of misinformation.

> *"A well-informed public is central to the proper functioning of a modern democracy ."*

# 1 Introduction

A well-informed public is central to the proper functioning of a modern democracy (Milner, 2002), like Canada. However, recent data indicates that 59% of Canadians are concerned about the integrity and truth of information that they are exposed to online and that 23% of Canadians use social media as their main source of news (Statistics Canada, 2023, 2024). In one survey, 79% of respondents indicated that steps should be taken to reduce fake news online, and many people have changed their news consumption habits (Fedeli, 2019). This concerning trend of the online spread of false information, or misinformation, has been the subject of recent research and debate regarding consequences for individuals, society, organisations, and governments.

## 1.1 Definitions

To better understand misinformation and cybersecurity, it is important to first clarify the meaning of key terms as they will be used in this report to avoid potential confusion.

## 1.2 Misinformation

Misinformation is defined as information, which is untrue, exaggerated, inaccurate, misleading, deceptive, confusing, manipulative, erroneous, or unverified, and is disseminated through traditional or digital means without deliberate or malicious intent (Aïmeur et al., 2023; Canadian Centre for Cyber Security, 2024; OECD, 2022b; Santos-D'amorim & Miranda, 2021).

## 1.3 Disinformation

Disinformation can also be defined as false, exaggerated, inaccurate, misleading, deceptive, confusing, manipulative, or erroneous information, but is specifically created or disseminated through traditional or digital means to intentionally cause damage and/or harm to any person, corporate entity or

public sector organisation for economic, financial, or political purposes (Aïmeur et al., 2023; Canadian Centre for Cyber Security, 2024; OECD, 2022b; Santos-D'amorim & Miranda, 2021).

## 1.4 Malinformation

Malinformation is defined as compromising, sensitive, damaging information that is true but deliberately used for harmful or deceptive purposes against a person, a group, an organisation or a country (Ireton & Posetti, 2018; Santos-D'amorim & Miranda, 2021; Walker, 2019).

Cybersecurity is the strategic application of technologies, processes, and human practices to protect systems, networks, programs, devices, and data from cyber-attacks. It encompasses the prevention, detection, and response to threats through the implementation of tools, policies, security safeguards, training, and best practices. It is a multifaceted approach that aims to mitigate risks and protect a cyber environment, including organisational and user assets such as computing devices, personnel, infrastructure, and information, ensuring their security and resilience against unauthorized exploitation (Global Knowledge, 2024; IT Governance, 2024; ITU, 2024; National Institute of Standards and Technology, 2018).

## 1.5 Development of Misinformation

While the spread of misinformation is not a new phenomenon, it has become increasingly problematic (Piccolo et al., 2019). The infiltration of communications technology, particularly social media, and the increased interconnectedness of our everyday online lives has encouraged the dissemination of

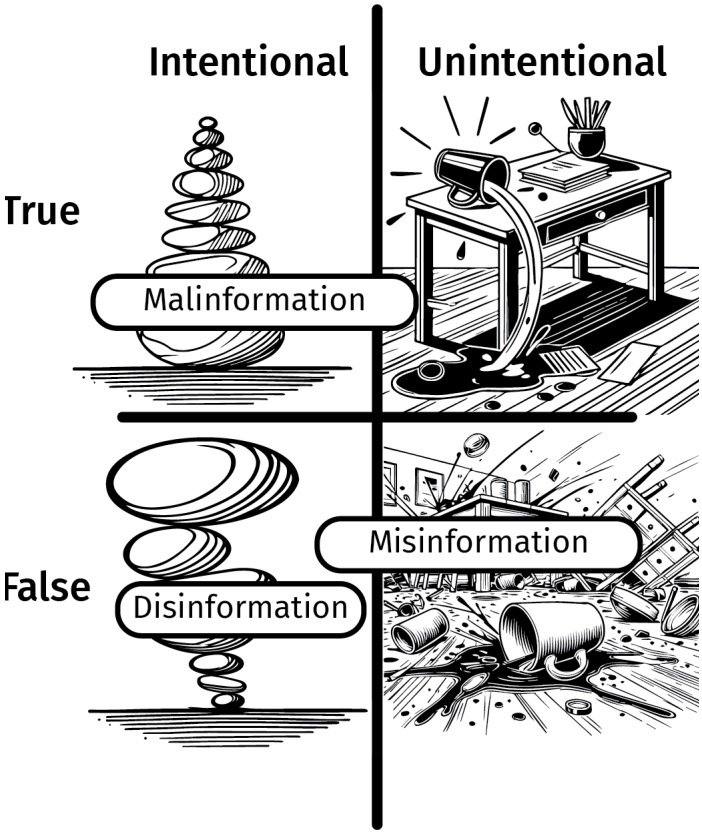


*Figure 1: Categorisation of information threat types*

misinformation (Suarez-Lledo & Alvarez-Galvez, 2021). The use of social media as a source of news, and to share news articles (whether true or false) has increased in recent years, particularly among older social media users (Grinberg et al., 2019a; Moretto et al., 2022). Several key elements of social media, including the ease of use, availability, speed of information diffusion, and difficulty in verifying information, create a breeding ground for misinformation (Thai et al., 2016). While this sharing of inaccuracies might seem harmless, it is important to note that the intentional spread of harmful disinformation by fake profiles and bots on social media can in turn develop into the unintentional spread of misinformation by real social media users (Guess & Lyons, 2020). The intermingling of truths, harmless inaccuracies and harmful untruths creates a modern information environment that is difficult to navigate.

Some of the first instances of social media

being used to spread misinformation and disinformation were in 2014, in the case of Russia-based 'troll armies' being employed to flood online forums with anti-Western and pro-Kremlin comments (Posetti & Matthews, 2018). In concerning later cases, analysis of social media posts in 2016 identified the use of fake profiles on Facebook and bots on X (formerly Twitter) during the UK Brexit referendum and later US Presidential election to post anti-remain and anti-Clinton messages (Posetti & Matthews, 2018). The trend of increasing disinformation online has only continued to the present with stories surrounding the war in Ukraine being a particular focus for campaigns(Post, 2024) .

Besides politics and wartime propaganda, online misinformation has also been observed to decrease public trust in governments and health agencies with harmful consequences. Research has demonstrated how the spread of online misinformation during the COVID-19 pandemic prevented effective control of the virus and successful implementation of vaccines. Reports emerged of people drinking disinfectant, methanol, and alcohol-based hand sanitizer based on misinformation spread online about how to combat COVID-19 (Islam et al., 2020). Rumours about the safety, efficacy, and intentions behind COVID-19 vaccines also hindered efforts to achieve widespread vaccine coverage (Lee et al., 2022).

The popularisation of large language models (LLMs), generative AI, and deepfakes can undoubtedly impact the dissemination of misinformation. By creating realistic but fabricated content, generative AI can be used to create fake news, misinformation, and propaganda in text, image, and video format that can be shared widely and easily on social media platforms (De Angelis et al., 2023).

While the dissemination of misinformation may not be a new phenomenon, the connected nature of modern life has seen it become a constant in the everyday lives of the public. It has become both pervasive and perverse, presenting a source of harm to individuals and institutions alike. As technologies continue to develop, it is likely that this trend will continue, making misinformation an important problem.

## 1.6 Purpose of the Report

This report aims to delineate factors relevant to misinformation, including political, regulatory, societal, and technological elements. The report also describes the cybersecurity issues posed by the spread of online misinformation and the potential impact of misinformation on society, as well as on public and private sector organisations. The report will finally propose key recommendations to reduce the spread and impact of online misinformation.

To prepare this report, peer-reviewed literature and reports relating to misinformation, disinformation, and cybersecurity were collected and analysed and combined with the outcome of discussions with topic experts as well as industry and public sector stakeholders.

Consequently, this report provides an overview of the current literature of misinformation, disinformation, and cybersecurity, and how these three elements interact. It provides a foundation for continued research and debate into the impact of misinformation on society, organisations, and governments, and how this relates to cybersecurity vulnerabilities.

# 2 Factors Relevant to Misinformation and Cybersecurity

## 2.1 Political Factors

The dissemination of misinformation is significantly influenced by the domestic and international political environment. Political misinformation includes false information about political groups, processes, or figures, and contributes to political division, political mistrust, democratic backsliding, and can exacerbate societal polarization (Koetke et al., 2023; Kozyreva et al., 2020; Lewandowsky et al., 2012). Political misinformation is easily accessible and does not require expert knowledge of the political landscape to be convincing. It is often driven by the political climate, can disproportionately affect the disadvantaged and is exacerbated by legitimate political and special interest groups.

Extreme political attitudes, particularly among the far-right, are disproportionately associated with the creation and spread of online misinformation, emphasizing the role of political ideologies in shaping misinformation sharing behaviors (DeVerna, 2024; Pretus et al., 2023). Further, the impact of these groups may be augmented by other factors. Older individuals and those who are more politically conservative tend to consume, believe, and share more political misinformation online, underscoring the influence of age and political orientation on the dissemination of false information(Grinberg et al., 2019b; Mosleh & Rand, 2021; Scheufele & Krause, 2019). For example, a report from the Louisiana State University found that of the known fake news stories that appeared in the three months before the election, those favoring Trump were shared a

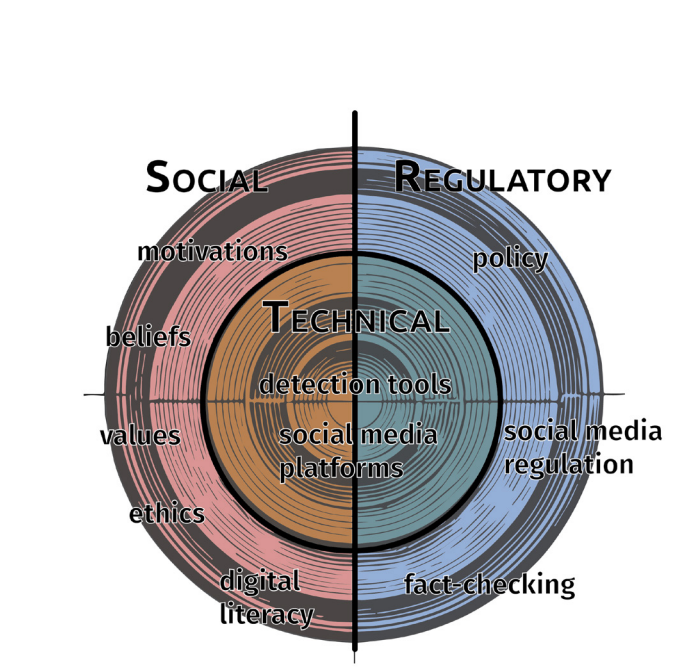


*Figure 2 "Social, regulatory, and technical aspects of misinformation adapted from Challenging Misinformation: Exploring Limits and Approach (Piccolo et al., 2019).*

total of 30 million times on Facebook, while those favoring Clinton were shared 8 million times (Georgacopoulos & Mores, 2020). The political climate may also be a factor, with political devotion and the desire to express a clear political stance are key drivers of sharing partisan misinformation (K. Zhou et al., 2024). The amplification of misinformation and conspiracy theories is often observed in politically polarized environments, where political leaders and media outlets take polarizing stances, contributing to the dissemination of false information (Su & Agyingi, 2024). The nature of this relationship may be complex, however, as misinformation may both increase polarization and be disseminated more due to cognitive biases. Misinformation is more likely to be consumed as factual and further disseminated when it confirms the readers' preconceptions (Sikder et al., 2020; Y. Zhou & Shen, 2022). In a polarized environment, this may result in misinformation that more clear-

ly confirms political values being accepted more easily and consequently being shared more often. Furthermore, the intertwining of misinformation and trolling - antagonistic and provoking posts and comments (Ortiz, 2020) - particularly in contentious and highly politicized contexts, reinforces each other and contributes to the spread of false information (Shah et al., 2024).

Politicians and interest groups are increasingly using misinformation as a strategic tool to influence public opinion and policy decisions. The spread of misinformation by political elites has been observed in various contexts, including social media platforms and public discourse (Mosleh & Rand, 2021; Pretus et al., 2023). This is particularly true during critical events such as elections and public health crises. Coordinated misinformation campaigns orchestrated by public figures and organisations often arise during election times, such as the claims of widespread fraud in the 2020 U.S. Presidential Election made by Republican politicians (Mosleh & Rand, 2021). These campaigns often target specific groups or individuals to sway opinions, influence policy outcomes, and mobilize support for particular agendas. For example, female politicians experience gendered disinformation on the basis of their identity as women, more than men do. Rather than attacking the policy decisions that women make, gendered misinformation plays on gender stereotypical characteristics and physical appearances to challenge female politicians (Di Meco & Wilfore, 2021). Gender stereotypes and biases spread through misinformed texts and photos have grave consequences on women's physical and digital security and have the potential to influence voters' views. In a 2016 global survey on female parliamentarians, the results demonstrated that 41.8% of respondents had seen "extremely humiliating or sexually charged" images of themselves shared on the internet (Inter-Parliamentary Union, 2016, p.4).

While legitimate political groups would be expected to operate ethically, there are concerns over the safeguards in place to ensure this is the case. In Canada, while the private sector and government institutions must adhere to Canada's privacy laws (i.e. Bill C-47), federal parties are exempt (Boutilier, 2023). Essentially, this means that there are currently almost no rules and zero oversight into how federal political parties collect, store, and use sensitive personal information about Canadian voters. So long as parties follow their own self-policed privacy policies, they can collect, use, disclose, retain and dispose of Canadians' personal information however they wish (Boutilier, 2023). By strategically using misinformation, political elites have the potential to exert considerable influence on public opinion and public behaviour, and shape perceptions about public policy on key issues. Even more worrisome, because there is no independent oversight, political parties have no legal obligation to notify Canadians if they experience a hack or data breach that compromises personal information (Boutilier, 2023).

Foreign actors have an interest in disseminating misinformation in their attempt to meddle in the internal politics of another nation. The use of social media platforms to disseminate misinformation and influence voting behavior has been a common strategy employed by foreign entities seeking to interfere in elections (Eady et al., 2023). The proliferation of misinformation by foreign entities can undermine trust in democratic processes and foster belief in falsehoods, ultimately affecting public perceptions and behaviors within the target state (Goldsmith & Horiuchi, 2023). For example, in the United States, the Congress and the FBI's investiga-

tions concluded that the spread of fake news – fabricated information that disseminate deceptive content, or grossly distort actual news reports, shared on social media platforms – during the 2016 US presidential election was created and disseminated to influence the election. These kinds of examples provide evidence of various actors within the domestic and the international realm with an interest in producing and disseminating misinformation to advance their political goals based on inaccurate or misleading premises.

## 2.2 Policy & Regulatory Factors

Debates are ongoing about the strategies that governments and private sectors should use to tackle misinformation or disinformation, especially deepfakes as a form of technology-generated deceptive information. It can also be difficult to discern the ill-intended creators of disinformation from those who naively share misinformation online. The spread of misinformation without the intent to cause harm or deception is not a criminal offence in Canada but can have far-reaching consequences. Nonetheless, the intent to commit an act is an important element in the design and implementation of a law, and without it a law focused solely on an act may be considered unjust, inapt or reduce the likelihood of conviction by jury, particularly where the offense is considered minor (White & Roberts, 1985) Therefore, legislation and policy should focus on the creation and distribution of disinformation. In other words, it should target false information spread with the intent to cause harm or to deceive.

A key concern for the success of these efforts is to determine which stakeholders are to be held accountable and how to achieve this. An initial hurdle is that disinformation and misinformation can be spread online without being subject to geographical or jurisdictional limitations. Often, "sources may be anonymous/pseudonymous and dispersed (including across platforms and borders), making it difficult, if not impossible, for any one nation-state or platform to effectively address it in isolation" (Pielemeier, 2020, p. 921). We have to note that a disinformation/misinformation creator or circulating digital platform may be geographically out of reach of specific in-country laws and regulations, which is a possible factor for the ineffectiveness of certain disinformation or misinformation-focused regulations (Chesney & Citron, 2018). In addition, it has been argued that technology enterprises, such as Facebook, are unlikely to self-regulate misinformation and disinformation on their own platforms (Waldman, 2018). It is therefore important for various stakeholders across cross-sectoral industries to bear the responsibility for their specific role in the spread and impact of misinformation, which presents a challenge in the implementation of policy and regulation. The OECD also highlighted that "efforts to curb mis- and disinformation must also be considered hand in hand with the full preservation of free speech" (OECD, 2022a), which is why any new regulations adopted should be with an acceptable consensus between key stakeholders.

A unified approach to creating and implementing legislation is necessary to encourage technology companies to take an active role in the reduction of harmful disinformation on their platforms. The Online Safety Act (2023) recently passed in the UK holds social media companies accountable for being transparent about the kinds of potentially harmful content they allow, and to give their users more control over the types of content they are exposed to. This legislation also criminalizes the creation and spread of disinformation; namely "sending false information intended

to cause non-trivial harm" (Online Safety Act, 2023). Crucially, companies can be fined, and criminal action taken against senior managers who fail to ensure their companies are adhering to the requirements set out in the Online Safety Act (2023). In addition, the Act gives powers to act against companies based outside of the UK, provided they have a significant number of UK-based users or relevant links to the UK.

## 2.3 Economic & Financial Factors

Disinformation spread can be driven by profit-seeking behaviours and due to financial incentives (Diaz Ruiz, 2023). Highlighted by Di Domenico & Ding (2023), the spread of misinformation can influence people's behaviors, including their financial decisions. Incidents on social media can have significant economic implications for companies, for instance on X (formerly Twitter) where a fake account using Eli Lilly's brand name posted that insulin would be given away for free resulting in a 4.37% drop in the brand's stock price, or when the sports brand Balance received considerable backlash after misinformation that the brand was associated with far-right movements spread online (Di Domenico & Ding, 2023). In addition, fake negative customer reviews can be classed as direct misinformation, impairing brand reputations and eroding customer trust, and on a larger scale can reduce public trust in the marketplace (Di Domenico & Ding, 2023).

It has been stipulated that the online advertising market by its designs that monetizes expansive engagement, i.e., promoting the active participation of new consumers, incentives the creation of content designed to 'go viral', including that which does so by means of circulating controversial claims, adversarial narratives, and deceptive content (Diaz Ruiz, 2023). Furthermore, misinformation is believed to cause instability in the market, as "the health of stock markets is dependent on the accuracy, timeliness, and transparency of information" (Cheng et al., 2023, p. 1). As an example, in 2022 Cointelegraph made a misleading online post regarding an operation involving the Securities and Exchange Committee (SEC) and the Exchange-Traded Fund (ETF) as two important intermediaries in the financial market. The incident was believed to cause a brief 5% spike in bitcoin's price and cause big losses for traders (Ozair, 2023). Lack of frameworks to demonetize certain false or misleading online contents (European Commission, 2022) constitutes one of the key weaknesses for stakeholders to address. It has been found that false information can, among other things, have an impact on financial markets and mislead individuals and companies in their financial decisions, which drives market inefficiency (Fong, 2021).

## 2.4 Societal Factors

The rise of the Internet and the widespread use of social media platforms have fundamentally changed how information is consumed and shared, which may have contributed to the current environment allowing the rapid dissemination of misinformation (Borges do Nascimento et al., 2022). The usability of social media platforms plays a crucial role in disseminating information, both accurate and misleading. In fact, research demonstrates that false news often spreads faster than real news online (Harrison, 2024). Various factors exacerbate the spread of misinformation on social media, including social media fatigue, cognitive biases, and narcissism, which can influence the likelihood of an individual to share false information (Bryanov & Vziatysheva, 2021).

The nature of social media platforms themselves is also an important contributor to the rapid dissemination of misinformation online. Social media platforms are structured to enable the rapid and widespread sharing of information, making them ideal channels for the propagation of both accurate and false information (Ahmad & Murad, 2020; Suarez-Lledo & Alvarez-Galvez, 2021). These platforms, designed to connect people instantly, enable misinformation to rapidly spread and reach a wide audience, significantly amplifying its impact (Muhammed & Mathew, 2022). Users do not always have the ability to retract or permanently delete their posts, leaving a permanent digital footprint that may not reflect the current opinions of the poster. The flexibility and segmentation provided by social media platforms enable individuals to passively select information that confirms their existing beliefs, creating echo chambers that reinforce biases and limit exposure to diverse viewpoints (Chuai & Zhao, 2022; Pantazi et al., 2022). If people are given news that conflicts with their previously held understandings (i.e. their mental models) it can induce a cognitive dissonance, or a discomfort caused by the attempt to hold contradictory beliefs or values (Konstantinou et al., 2019). For this reason, social media users are less likely to pay attention to anything that does not align with their personal values or beliefs, leading them to engage more with content they find cognitively comfortable and consequently informing the media selection systems of the platform.

This is also exacerbated by confirmation bias, or the tendency for humans, where they see what they expect or want to see, to be less critical, thus being unlikely to fact-check (Konstantinou et al., 2019). Consequently, social media users may be more inclined to uncritically accept recommendations on social media platforms that amplify their existing beliefs, leading the platform to make more of such recommendations. The resulting repeated exposure to similar misinformation can increase trust in false beliefs through the illusory truth effect, which is that the repeated exposure to information results in it being perceived as more truthful (Ahmed & Rasul, 2023). Pennycook et al. (2018) demonstrated that individuals are more likely to deem a false statement true the more times they are exposed to it. This may be because the information is recognized as familiar, even though the reader can't necessarily remember where or in what context they encountered it before.

The problematic belief in false information is compounded by individuals then sharing this information with their social network, with this behaviour implicitly suggesting an endorsement of the content. Social proof, a concept that suggests that people use the actions of others to define correct behaviour, further amplifies the effect. That is, people seeing others around them sharing a particular narrative may set such acts as normal and expected. Further exacerbating the issue is the tendency for people to share more extreme content more often. When exposed to information which is threatening, or in alignment with their fears, individuals are more likely to share the information, this further speeds up the spread of this information (Oh et al., 2013). More engagement with misinformation adds to its credibility and its spread (Avram et al., 2020; Chuai & Zhao, 2022).

The interaction between social media platforms and human cognitive biases creates a problematic scenario. Frictionless sharing allows for individuals to share information easily and mass personalization (see 3.5.1 below) selects engaging content passively based on their behaviour. Human cognitive biases result in information being shared and endorsed uncritically and media preferences being set

based on comfort and popular narratives within a social group. This results in an online environment where misinformation can be widely shared without being challenged by critical examination or contradictory messages.

### 2.4.1 Undermining Public Trust in Cybersecurity

Misinformation within the realm of cybersecurity has damaged public trust by diminishing confidence in digital systems and data protection. Exposure to false or misleading information about cybersecurity practices, threats, or incidents can decrease trust in online platforms, government entities, and information sources. Cyberattacks like phishing, ransomware, and distributed denial of service (DDoS) not only cause direct effects such as data breaches and financial losses but also contribute to the deterioration of public trust in online platforms and e-government services (Al-Hawamleh, 2024). Data breach incidents can be argued to have severe consequences, including the undermining public trust in important institutions (Banner, 2022). A survey conducted for the Office of the Privacy Commissioner of Canada in 2013 revealed that only 21% of Canadians believed the federal government took its privacy-related responsibilities very seriously (Office of the Privacy Commissioner of Canada, 2013).

In the context of cybersecurity, misinformation can generate confusion, fear, and uncertainty among the public, reducing their confidence in the security of their personal data and online systems and services (Arambul et al., 2023). For example, ransomware groups are intentionally using misinformation to increase their social status and to trick authorities. They make false claims about ransomware and data breaches and rely on others, including cybersecurity experts, to spread these false claims (Bracken, 2024). Moreover, spreading false information during a cybersecurity incident can worsen the attack's impact and erode public trust in organisations' ability to protect sensitive information. During crises, malicious actors can exploit public trust vulnerabilities by disseminating false information about security protocols, data breaches, or the effectiveness of cybersecurity tools, thereby creating opportunities for cyber threats to succeed (Joseph et al., 2022). This intentional spread of misinformation can result in skepticism, confusion, and a lack of confidence in the cybersecurity measures implemented by organisations and governments, making it challenging to distinguish them from malicious actors.

## 2.5 Technological Factors

### 2.5.1 Mass Personalization and Social Media

The extent to which social media companies allow and promote account customization has a significant impact on the spread of misinformation and the formation of echo chambers (Barberá et al., 2015; Bruns, 2017; Cinelli et al., 2021; Dubois & Blank, 2018). Account customization is often achieved by means of mass personalization technologies that ease the configuration burden on consumers by leveraging algorithmic processes that operate in the background, computing big datasets to predict the preferences of each individual using the platform (Kotras, 2020). The customization of preferences based not on informed and deliberate choice but by behaviour can have undesirable consequences, such as informational echo chambers.

Echo chambers are environments where individuals are constantly and repeatedly ex-

posed to information which aligns with their pre-existing beliefs. Echo chambers are created and reinforced through the algorithms prevalent in social media, which are based on quantitative values like likes and shares. The algorithms lead to the creation of silos, which isolate users from perspectives contrary to their own (Avram et al., 2020; Oh et al., 2013). Research suggests that people develop more and more extreme beliefs due to the reinforcement they receive from echo chambers (Sunstein, 1999). These echo chambers can be found in blogs, forums, and social media websites (Barberá et al., 2015; Edwards, 2013; Gilbert et al., 2009; Grömping, 2014; Quattrociocchi et al., 2016). Modern technology that only allows individuals to have exposure to this selective information can lead to increased polarization, damaged critical thinking and the elimination of productive dialogue (Sunstein, 1999).

While the operation of platforms is problematic, there is also the potential for them to be a part of the solution. While platforms may be reticent to alter the underlying technologies that promote engagement with platforms, arguably a feature important for the success of social media platforms, they do not preclude direct actions to curb misinformation. Social media companies are capable of taking such actions and have done so. For example, Meta took down hundreds of groups, pages and accounts on Facebook and Instagram due to misinformation, defined by their head of cyber security policy as "co-ordinated inauthentic behaviour" (Delhi, 2019). These accounts posted "massive amounts" of content repeatedly or were fake accounts, and were removed due to suspicious activity, not due to the content itself (Delhi, 2019). WhatsApp, also owned by Facebook, limited users forwarding messages to five times, down from 20, after angry villagers were incited to mob lynchings via WhatsApp messages (Delhi, 2019). This indicates that it is technically possible for social media operators to intervene.

### 2.5.2 Deepfakes

Deepfakes are a form of manipulated graphical (image and video based) content, where a face and/or body parts are puppeted by another using advanced AI techniques(Nguyen et al., 2019). The resulting realistic images and videos are often employed to deceive viewers, spread misinformation, and manipulate public opinion. This presents a threat to the integrity of information, and to cybersecurity as a whole.

There are technical countermeasures that have been developed in order to allow for the computational detection of deepfakes. These make use of various methods:

1. Analyzing visual artifacts relies on the identification of inconsistencies in visual aspects such as lighting, shadows, or texture of the video (Matern et al., 2019). The visual aspects can make evidence of tampering more evident.
2. Blending boundaries focuses on detecting irregularities at the edges where AI generated content merges with real content (Li et al., 2020). The irregularities are evident due to imperfect integration of the content.
3. Mouth movements is a method that analyses speech patterns to see if the lip synchronization matches natural human behavior (Haliassos et al., 2021).
4. Behavioural biometrics assesses unique patterns in how individuals move or express themselves (Agarwal et al., 2020; Nadimpalli & Rattani, 2022).

Even with these resources, however, detecting deepfakes remains a challenging task. The identification of deepfakes by technical means has been difficult due to the constant advancements in this form of technology (Aghajari et al., 2023). High quality and

realistic images and videos can be created by computational models which are able to produce output which resembles original data (Nguyen et al., 2019; Zhang et al., 2017). Furthermore, aspects such as poses, facial expressions, and lighting can be preserved when faces are swapped (Korshunova et al., 2017). This makes it incredibly difficult for forensic computational models to differentiate deepfakes from reality.Deepfake images and videos represent a potential mechanism for the transfer of misinformation that is difficult to combat via technical means as the technology being developed to perfect the generated imagery appears to be currently leading the detection technologies.

### 2.5.3 AI Generated Content

Artificial Intelligence-Generated Content involves using generative AI algorithms to either assist or replace human effort in producing unique content (Wang et al., 2023). This process is driven by user inputs or specific requirements, enabling faster content creation at reduced costs compared to traditional methods (Wang et al., 2023). This content can be in almost any digitally transmissible form, including text, audio, images and video.

One major concern of AI-generated faces and poses is the risk it presents regarding security, identity verification, and misinformation detection (Khoo et al., 2021). Studies show that individuals are not able to tell the difference between real and AI-generated images. This puts them at the risk of being tricked into believing lies through images. Surveys have shown that 40% of respondents are unable to reliably tell the difference (Pocol et al., 2024). Public education regarding these topics is a very important step in tackling misinformation (Poredi et al., 2024).

The process of creating AI-generated content also suggests further legal and ethical issues. This is primarily because the AI is generally trained on data from the internet. It can therefore possibly be trained on source materials that are erroneous, fictitious or part of a disinformation campaign. As a result, all content generated using that dataset holds the risk of further advancing that misinformation. Because AI does not yet always cite its sources, there remains a concern for the legitimacy of any information it generates. Even when the information is correct, the authority of its truthfulness is difficult to measure. Furthermore. As the sources are not given credit, the sources are robbed of their intellectual property without consent. One such example is the legal action taken against GitHub's copilot which allegedly used licensed code without attribution (Butterick, 2021). Identifying the content that is used as training data can effectively tackle this issue. This approach is being employed by companies and platforms such as Stability AI and Spawning AI (Beaumont, 2022).

AI is, and will continue to be, used to disrupt society through misinformation, and addressing the authenticity of visual content is a difficult challenge to overcome (Goldstein et al., 2024; Poredi et al., 2023; Solomon, 2023; Solomon et al., 2022; Solomon & Cios, 2023). Addressing the issue of deepfakes and other AI-generated, misinformed content, is a moving target - as the datasets, methods, and technologies grow, the detection becomes increasingly difficult.

# 3 Countering Misinformation

Countermeasures must be taken in an effort to reduce the spread and impact of online misinformation and disinformation. The tables below provide an overview of technological and human-centric countermeasures to reduce misinformation, with both a proactive and reactive approach. A selection of countermeasures will be described in further detail below. By combining proactive measures with reactive responses, both online platforms and individuals can work towards minimizing the impact of misinformation and promoting a more informed digital environment.

## 3.1 Approaches

### 3.1.1 Reactive Approaches

Reactive approaches to countering online misinformation involve responding to false information after it has been disseminated.

| **Proactive Approaches** | |
|---|---|
| Automated Tools for Detection | Deploying automated tools to detect misinformation and provide notifications to users and moderators. |
| Watermarks | Using digital watermarks to verify the authenticity of content and prevent the spread of altered or fake information. |
| Algorithm Transparency | Ensuring that the algorithms used by social media platforms are transparent and understandable to the public to reduce the spread of misinformation. |
| Advanced Machine Learning Models | Developing sophisticated machine learning models to detect and mitigate misinformation more effectively. |
| Detection of Deepfakes | Implementing AI technologies to identify and block deepfake content. |
| **Reactive Approaches** | |
| Automated Content Labelling | Applying labels to existing content that has been identified as misinformation to warn users. |
| Bot Detection | Identifying and removing automated accounts (bots) that spread misinformation. |
| Algorithm Transparency | Making changes to algorithms to reduce the visibility of misinformation once it has been identified. |
| Identifying and Removing Disseminators of Misinformation | Tracking down and removing accounts and sources that are consistently spreading false information. |
| Advanced Machine Learning Models | Using advanced models to continuously monitor and react to new forms of misinformation. |
| Detection of Deepfakes | Employing AI to remove or label deepfake content that has already been shared. |

*Table 1- Misinformation Countermeasures - Technological Solutions*

| Proactive Approaches | |
|---|---|
| Prebunking | Educating individuals about common misinformation tactics before they encounter them to build resistance to false information. |
| Inoculation | Providing people with a weakened form of misinformation to help them build defenses against future exposure to stronger forms. |
| Improving Social Media Regulation | Enhancing policies and regulations for social media platforms to ensure they take responsibility for reducing the spread and impact of misinformation. |
| **Reactive Approaches** | |
| Debunking | Actively correcting false information that has already spread by providing factual and verified information. |
| Crowd-Generated Misinformation Flagging | Leveraging the public to flag and report false information, allowing for quicker identification and correction. |
| Fact-Checking | Utilizing fact-checkers to verify the accuracy of information and debunk false claims. |
| Positive Use of Shame | Encouraging social accountability by highlighting and shaming the spread of misinformation. |
| Friction Introduced in Social Media | Implementing measures that slow down the sharing of news on social media to give users time to verify information before spreading it. |

*Table 2 - Misinformation Countermeasures - Human-Centric Solutions*

While it does not effectively address the underlying causes of misinformation, a reactive approach may be necessary for removing misinformation in an immediate sense.

#### 3.1.1.1 Automated content labelling

Online platforms have increased the use of automation to label large volumes of content, while relying on automated interventions to moderate the credibility of information or sources uploaded under their user's names (Alaphilippe et al., 2019). Automated content labelling relies on machine learning to classify the content moderation process into problem categories, as well as to add into a database of known unwanted content, such as misinformation. While they can quickly label a post or a content as false information with versatility, it can struggle to stay consistent and reliable without human intelligence reviewing it (Roozenbeek et al., 2023). Therefore, significant risks of error moderation can pose a threat to counter measures of misinformation, as social media platforms tend to fail to provide the efficacy levels of their technological interventions (Roozenbeek et al., 2023).

#### 3.1.1.2 Individual and crowdsourced misinformation flagging

While professional fact-checkers and automated content labelling provide objective insights on factual information, they generally do not directly engage with misinformation spreaders on social media platforms (He et al., 2023). Ordinary social media users can play a crucial role when a platform allows them to report and flag misinformation posts (Micallef et al., 2020). An example is X's (formerly Twitter) Birdwatch, where users

can actively engage to identify posts, they believe to be misleading or false, although they do not allow user-to-user communication and countering misinformation directly on X. Unlike professional fact-checkers that indicate clearly if a content is true or false, social media platforms like Facebook display red flags on posts that lack credibility according to fact checkers (Aghajari et al., 2023). Although warnings against content are known to reduce its perceived accuracy, they can only signal the credibility as false and do not offer any additional information on the context. Furthermore, these incentives can also prompt people to click on the false content, therefore increasing its visibility and backfiring counter measures (Aghajari et al., 2023).

#### 3.1.1.3 Identifying and removing disseminators of misinformation

To limit the spread of false information from their platforms, social media companies can focus on a micro level at individual users who engage in more misinformation behaviors. Meta (formerly Facebook) and X are known to identify and possibly remove any account that exhibits “inauthentic behavior”, which is defined as the use of social media, such as a page, followers or links, to mislead people(Aghajari et al., 2023). Users who engage in coordinated inauthentic behaviour can have their accounts identified and disabled, with their content becoming inaccessible to others (Aghajari et al., 2023) .

#### 3.1.1.4 Debunking

Debunking refers to the presentation of verified information to establish that previous content was misinformation and addresses the psychological process of correcting the information (Chan et al., 2017). Debunking can be fact-based or logic-based, which focuses more on the manipulation techniques and the epistemic quality of the false information (Vraga et al., 2020). Meta uses debunking to moderate content on their platforms but has been criticized for lacking transparency over what kind of content gets limited on their social media (Roozenbeek et al., 2023). Therefore, the effectiveness of fact-checking depends in part on the cooperation of large online platform companies.

#### 3.1.1.5 Fact Checking

Fact-checking is a tool, refined through journalism, that checks the authenticity of statements (Graves et al., 2016). It is a proactive process in which claims are verified against available evidence or factual records before or shortly after they are published. It is distinct from debunking it that it is an element of rather than the complete process of correcting retained information. It is a rather popular reactive approach, as multiple fact-checking sites such as Snopes.com, FullFact.org and StopFake.org, are used to correct misinformation.

While fact-checking can reduce belief in false claims, it can also increase false beliefs through a misinterpretation of the facts (Nyhan et al., 2013; Nyhan & Reifler, 2015). Studies have shown that the emotions attached to misinformation affect the response people have in terms of corrections (Moore, 2016; Morgan et al., 2013). As a consequence, it is possible for efforts relying on fact checking to have unintended and undesirable outcomes.

Social networks are very important in fact checking. Studies have shown that fact-check articles are taken more seriously when shared by friends than they are when shared by strangers (Hui et al., 2018). With that said, however, it can be advantageous to exclude biographical information about fact-check authors to maximize effectiveness (Garrett et al., 2013). Fact checking needs to have a large reach as

it is generally slower than the spread of misinformation (Vosoughi et al., 2018). Research related to COVID-19 has suggested that younger, and more politically liberal users are more likely to fact-check online content ( Rich et al., 2020; Robertson et al., 2020).

#### 3.1.1.6 Bot/Automation Detection

A social media robot, or social bot, is an automated program designed to produce content and engage with users on social media, often mimicking human behavior (Gorwa & Guilbeault, 2020). It interacts in ways that resemble how a real person would act online, creating the appearance of genuine user activity. Social bots are typically public-facing and are designed to seamlessly blend into social media environments, automating responses, posts, and interactions to influence conversations or amplify content (Gorwa & Guilbeault, 2020) .

While these bots may be deployed for any purpose, that does include influencing political circumstances and spreading misinformation (Gorwa & Guilbeault, 2020). The presence of bots on social media introduces several challenges related to integrity. Twitter bots have historically interfered with election results (Deb et al., 2019; Ferrara, 2017), and disseminated misinformation (Cresci, 2020). Contemporary bot detection techniques use approaches like graph neural networks . This has helped tackle issues such as fraud, misinformation, and poor recommendations (Dou et al., 2020; Lu & Li, 2020; Varlamis et al., 2022; Wu et al., 2020).

Despite technological advancements, it remains difficult to distinguish between humans and bots. This is a hurdle for detection purposes, particularly where the bots are designed with the intention to misinform (Cresci et al., 2017; Grimme et al., 2018). By understanding and addressing the technological factors that facilitate bot-driven misinformation, cybersecurity efforts can better protect the integrity of online discourse and ensure a safer, more reliable information environment.

These same bots, however, can also be created to detect other bots. While that would infect the online space in the sense that it would mainly contain bots, the bots meant for fact checking can detect other bots and eventually protect the users from being misinformed.

### 3.1.2 Proactive Approaches

While corrective and reactive measures for identifying misinformation can be effective, they are not without flaws, particularly those concerning their accuracy, expense, and timeliness. Ideally, preventing misinformation from finding traction in the first place would allow for safer practices online. Proactive measures can be used to counter false information, based on the principles of the inoculation theory (Lewandowsky & Van Der Linden, 2021). Inoculation rests on the baseline that if individuals are forewarned that they might be exposed to misinformation online and exposed to weakened examples of the ways in which they might be misled, they will develop an immunity to misinformation. Under a biomedical analogy of a vaccination for brainwashing, it would be considered to be spread throughout the cyberspace like a viral pathogen. If we pre-emptively expose a community to misinformation, then an immunity to the viral persuasion could be

developed (Pilditch et al., 2022). Inoculation theory has been successfully applied to misinformation about climate change, conspiracy theories, hate speech online and COVID-19 (Lewandowsky & Van Der Linden, 2021).

Pre-emptive approaches, such as prebunking are well-tested and are grounded in a robust evidence base dating back to the 1960s, proving effective across a wide range of scenarios (Harjani et al., 2022). By proactively addressing misleading narratives or techniques across various topics and domains, it scales more efficiently than combating individual claims. By virtue of its pre-emptive nature it is able to take a non-accusatory and non-judgmental tone and consequently can foster openness and encourage audiences to engage with preventative interventions. Furthermore, it can remain apolitical by focusing on misleading techniques rather than specific claims, demonstrating effectiveness across the political spectrum, including among individuals with conspiratorial beliefs (Harjani et al., 2022).

#### 3.1.2.1 Prebunking

Inoculation theory also includes prebunking, which can be both an active and passive technique to build pre-emptive resilience to misinformation. Prebunking focuses on educating how people are commonly manipulated and misled online, rather than directly challenging misinformation or automatically categorizing content as being true or false (Harjani et al., 2022). This psychological technique can effectively inoculate individuals against misinformation they did not see before. It seems to be the most effective when people can generate their own counterarguments, which boost their confidence in their own truth-discernment abilities and reduce self-reported willingness to share misinformation on social media (Omoregie, 2021). Active technique-based inoculations can be found in the form of online games, such as Bad News, Go Viral!, Radicalise, and Harmony Square, which exhibit a social media-like environment to improve people's ability to recognize misinformation (Pilditch et al., 2022). For example, Bad News simulates a social media feed where players are impersonating a fake news producer who tries to gain as many followers as quickly as possible, all that while trying to not lose credibility (Roozenbeek & Van Der Linden, 2019). The purpose of Bad News is to enable individuals against the known methods of misinformation by allowing them to generate their own fake news (Lewandowsky & Van Der Linden, 2021).

While optimistic, improving people's skills to identify false and misleading content does not always lead to changing people's intention to share misinformation. For example, many people may know that a piece of content is untrue and still spread it anyway motivated by political or social reasonings (Aghajari et al., 2023). It is also difficult to prebunk individuals against all of the misinformation online (Pilditch et al., 2022). To improve the efficacy of diminishing the spread of misinformation, improving individuals' media literacy skills through critical thinking skills can help increase skepticism towards false information (Badrinathan, 2021).

#### 3.1.2.2 Blockchain Technology

Blockchain technology is a decentralized record keeping system that multiple parties have access to (Portmann, 2018), that stores data

in a manner such that it is accessed securely and transparently (Nakamoto, 2017). Each participant in the record keeping network generates a cryptographic code, ensuring the system remains secure and trustworthy. Because of this, blockchain is an ideal technology when it comes to data integrity and authenticity (Vardhan et al., 2023).

Blockchain is known for its ability to detect authenticity. When used with the right complementary technologies, it can track and verify goods (in this case, information) throughout a supply chain (Kshetri, 2018). With the help of digital tokens and smart contracts, blockchain creates a unique identity for each product (Zheng et al., 2017). This ensures the authenticity of every product (information) and eliminates counterfeits (misinformation), enhancing the trust and reliability goods at every stage, from production to distribution, enhancing trust in the supply chain (Vardhan et al., 2023).

Smart contracts on the blockchain can automate and secure the verification of digital content (Dai et al., 2019). These contracts work using predefined conditions and can ensure that the content has not been tampered with (Guru, 2024). The way in which blockchain can be used is to integrate a watermarking on digital content (Abrar et al., 2021). By encrypting watermark information and storing it in the blockchain, the authenticity of digital images can be ensured (Huang & Yi, 2023).

Despite the advantages, however, the implementation of blockchain for digital content verification presents challenges. One issue is the limited storage on blockchain (Indapwar, 2020). Solutions such as the Interplanetary File System (IPFS) provide the necessary infrastructure for decentralized content storage, complementing blockchain's capabilities by ensuring that no central entity controls the data (Golosova & Romānovs, 2018). IPFS also breaks the data into smaller chunks making it more resistant to censorship and ensuring its integrity (Badari & Chaudhury, 2021; Benet, 2014; SK et al., 2024).

While blockchain is secure and transparent, it can also lead to privacy issues (Heo et al., 2020). The transparency of blockchain means that all participants can access the records; this can include sensitive information (Feng et al., 2019). To address these concerns, advanced algorithms and techniques can be employed to enhance privacy and security (Heo et al., 2021).

Blockchain has historically been applied for the verification and the authenticity of diplomas and certificates. The implementation of a blockchain-based system at Al-Zaytoonah University in Jordan is an example that educational credentials are authentic and tamper-proof (Chaniago et al., 2021). This system uses student National IDs and smart contracts to verify diplomas, demonstrating blockchain's potential in various fields beyond supply chain management (Kanan et al., 2019). While blockchain is appreciated for its decentralized nature, its implementation often involves a centralized component. Federated blockchains, which are a type of blockchain that is not accessible to the public, are more practical for certain applications (Hoffman et al., 2020; Zheng et al., 2018).

It is important to note that watermarks are only effective under very specific circumstances. Firstly, the consumers have to take on the responsibility of verifying the watermark and its legitimacy (as it would likely often be forged) and secondly, there would need to be a consensus on what watermark would be considered legitimate; without a consensus, malicious third parties are incentivized to create their own watermarks to

continue the consumption of misinformation.

# 4 Factors Influencing the Adoption and Effectiveness of Misinformation Countermeasures

Various factors can reduce the effectiveness of misinformation countermeasures. As previously discussed, research has demonstrated that prebunking online misinformation can be effective at reducing the spread and impact of misinformation (Roozenbeek et al., 2023). Conversely, the process of passive prebunking may be limited in its effectiveness. In this approach, individuals are simply given a counterargument without further engagement. Research into prebunking and counterfactual thinking is also limited in its ability to reflect real-life situations. Social media use in real-life is varied and influenced by environmental factors. These include communication with others and the simultaneous use of multiple apps. Thus, prebunking countermeasures to misinformation may be hard to implement on social media in reality, and their real-world effectiveness may not match the potential expected based on research performed under lab conditions.

In addition, prebunking and/or inoculation messages may only be effective for certain individuals. In the context of COVID-19 misinformation, Amazeen and colleagues (2022) identified that specific inoculation messages had a detrimental effect on individuals who held pre-existing unhealthy attitudes regarding COVID-19 vaccines, increasing the likelihood that these individuals would believe the misinformation. While individuals who held healthy attitudes towards vaccines were likely to uphold these attitudes if the misinformation they are exposed to is prebunked, it is arguable that these individuals are not the true target of misinformation. Prebunking methodologies may also differ in their efficacy. For instance, logic-focused corrections, which undermine the rhetoric presented by the misinformation, are effective whether presented before or after the misinformation as opposed to fact-focused corrections, which provide accurate information to refute the misinformation, only reduce misperceptions when they occur after the misinformation (Vraga et al., 2020). Ultimately, prebunking countermeasures to online misinformation are limited in their effectiveness and logistical practicality.

Some traditional countermeasures like warning labels and factchecking have been largely ineffective in combating misinformation on social media platforms. Research has highlighted a few potential reasons for this, including that (mis)information sharing is a behaviour determined by social processes and that social media platforms foster an environment which promotes this (Jones et al., 2023). The impact of social cues intertwined within social media platforms, for instance, social reinforcement through liking or commenting on a post, has not been empirically assessed in the context of misinformation countermeasures. While it is unlikely that there will be a fundamental redesign of social media to remove these elements, the adaptation of certain features could discourage the spread of misinformation. “Taiwan,” for example, an online discussion forum for citizens of Taiwan to discuss proposals and share information, doesn’t allow replies to comments. The notion behind this was to deter trolls and discourage divisive conversation. In addition, a visual mapping of posts based on upvotes and downvotes was effective in creating like-mind-

ed groups and establishing gaps. Users then posted comments to try and win votes from both sides of the divide, gradually eliminating the gaps and instead promoting a unified response (Horton, 2018). This example demonstrates the potential advantages of adapting social media platforms and how this may reduce the spread and impact of misinformation. Despite these social benefits of these adaptations, the process of changing could incur a financial burden and may impact the performance of the platforms, making independent adoption of such changes unlikely. As a result, regulation may be necessary to reform the dangerous aspects of these platforms.

Countermeasures against misinformation require a multifaceted approach. This includes the development of legal policies, awareness campaigns, enhancing true content in mass media, and improving digital literacy among the population. Unfortunately, the effectiveness of these countermeasures may be limited due to the complex and evolving nature of misinformation dissemination and the challenges in reaching and educating diverse audiences (Borges do Nascimento et al., 2022). Furthermore, education alone may be insufficient to combat the impact of online misinformation and certain approaches to increasing public knowledge of misinformation may be counter-effective. Educational initiatives require individuals to be motivated to engage and can exclude those with low digital literacy, who may be the most likely to share misinformation in the first instance (Bronstein & Vinogradov, 2021). In addition, educational interventions do not eliminate vulnerability to misinformation, with one study demonstrating that 20% of individuals still believed misinformation even after being taught strategies to spot misinformation ( Guess et al., 2020).

In summary, countering misinformation requires a multifaceted and adaptable approach, as no single method, whether prebunking, fact-checking, or educational initiatives, can fully address the complex and evolving nature of misinformation. Social media's varied use and reinforcement of social behaviors complicate the effectiveness of countermeasures, while the psychological resistance among certain groups and the challenge of reaching individuals with lower digital literacy further limit the impact of these interventions. Regulatory reform, platform redesigns, and broad-based digital literacy efforts are critical, but these too face practical barriers.

# 5 Recommendations

## 5.1 Mandate Regulatory and Policy Changes

Implementing stringent policies that hold social media platforms accountable for reducing the spread of misinformation and disinformation is crucial. Such regulation has already been adopted in other countries. The UK Online Safety Act (2023) require social media platforms to monitor actively and fact-check content. The Act aims to protect users by holding social media platforms accountable to being transparent about which kinds of potentially harmful content they allow and present users more control over the types of content they want to see. The Act also criminalizes the spread of disinformation. Crucially, the Act applies to companies based outside the UK, provided they have links to the UK. This includes if the service has a significant number of UK users, if the UK is a target market or it is capable of being accessed by UK

users and there is a material risk of significant harm to such users (Online Safety Act, 2023).

Adopting cooperative and extraterritorial regulations based on the needs of the various key stakeholders identified could increase the chance of misinformation laws and regulations to be effective. While the enforcement of such regulations presents challenges, international cooperation resulting in broadly harmonized rules provides an economic incentive beyond any punitive measures, as operating in compliance with global norms may be more attractive to shareholders.

Under an ethically oriented policy framework, platforms should be encouraged to develop mechanisms reducing the incentives creating and distributing misinformation and disinformation. Platforms have developed a number of tools for both incentivization and dis-incentivization such as demonetization, temporary bans, and permanent deplatforming. The transparent application of these tools to misinformation would likely have a positive impact, while allowing for the just application of the process to be independently verified. Such a policy is likely to encourage stakeholders to promote a more responsible distribution and use of information, whether from an individual, collective or corporate point of view.

## 5.2 Increase Public Awareness and Education

Public education campaigns play a vital role in combating misinformation. While individuals should not bear the sole responsibility for verifying information, increasing their awareness about misinformation can significantly reduce its spread.

Educating people on how to identify misinformation, understand its signs, and verify facts can empower them to make informed decisions. The use of active prebunking systems such as those engaging serious games to enable logic-focused critiques of messaging show promise in promoting the appropriate skills to effectively navigate misinformation.

## 5.3 Enhance Public Digital Literacy Initiatives

Digital literacy programs are essential in teaching individuals how to navigate online information critically. These programs could help users to differentiate credible sources from dubious ones, fostering a more discerning and informed public.

These programs should be implemented across demographic groups, starting in schools. Initiatives such as Google's digital safety program (https://beinternetawesome.withgoogle.com/en_us) which promotes a 'Share with Care' approach to children, have shown promising results in helping children to understand and be critical of information they are exposed to online (Jones et al., 2023). Another group vulnerable to misinformation is older adults, particularly those with low digital literacy. Programs aimed towards older internet users should be prioritised, as older individuals are more likely to share misinformation.

## 5.4 Improve Automated Content Labelling and Algorithm Transparency

Effective content labelling systems are critical in helping users process and understand online information. By drawing from the effectiveness of other warning systems, such as those for smoking and alcohol, content labels can be designed to attract and hold users' attention. Labels that are clear and informative can enhance users' comprehension and awareness of misinformation.

Current algorithms are designed to form silos and echo chambers. This is done keeping the marketing aspect in mind, to keep people engaged with content by only showing them what they are certain to like and agree with. The current systems appear to be optimized for engagement and do not consider the mental, intellectual or physical health of consumers. Systems that more deliberately expose users to content from various ideologies would more effectively allow them to explore various perspectives. The mechanisms for the selection of content should allow for more active consideration by users. While it may be the case that if the algorithm was entirely transparent, content creators and companies could effectively manipulate it, it would also empower individuals to have a better understanding of why they see what they see. A well-implemented algorithm would also be designed in a way that transparent in its operation for users. would not necessarily expose it to possible exploitation.

## 5.5 Utilize AI for Detecting Fake Content

Artificial intelligence has significant potential in identifying and mitigating fake content, including deepfakes. Investing in advanced AI technologies can enhance the reliability and safety of online information by detecting and removing false content efficiently. AI's capability to analyse vast amounts of data in real-time makes it a powerful tool in the fight against misinformation. While it may be the case that the current tide of technological development is favoring the development of deepfake technologies, this need not continue to be the case. Investment in the development of deepfake related technologies is based on the content generation capabilities of these tools, which provide clear and valuable advancements for the development of content creation tools for legitimate purposes. These uses do not require that the result be technically indistinguishable from genuine content. As a result, it may be possible that the fake content detection AI may overtake fake content generation. Were private industry to develop only forensic-ready generation tools, or generation tools that inherently identify their source, this would effectively reduce the utility of this form of tool for mis and disinformation purposes. Furthermore, this would also reduce the utility of these advances to those developing malicious adaptations of these technologies. Assuming that the continued development of detection capabilities by large platforms and government-funded research, such technologies could overtake that possible from illicit operators.

# 6 Conclusion

This report effectively presents an examination of the political, regulatory, societal, and technological factors driving the spread of misinformation and its impact on cybersecurity. The challenges associated with misinformation, such as the erosion of public trust and the manipulation of democratic processes, are complex and persistent, particularly in the ever-evolving landscape of social media.

Resolving the problem of misinformation in Canada presents a significant challenge. Various countermeasures have been explored to address this issue, such as prebunking, fact-checking, and automated detection systems. Prebunking involves proactively presenting counterarguments before misinformation spread in a manner that considers the psychological processes required, but expectations for its success should be tempered as its effectiveness can be diminished in the complex dynamics of real-world social media environments. Fact-checking, though essential, should be considered as a component of a successful program as it frequently encounters difficulties due to the emotional and social motivations behind the spread of misinformation, which factual corrections alone cannot address. Furthermore, advances in AI and deepfake technology continue to outpace detection mechanisms, making it increasingly difficult to identify and mitigate the spread of false content.

Addressing the problem of misinformation requires a multifaceted strategy that combines regulatory reforms, technological solutions, and educational initiatives. Regulatory interventions should prioritize transparency in algorithmic practices and hold platforms accountable for the information they promote. Meanwhile, public education campaigns must focus on improving digital literacy to help users critically assess the information they encounter online. In addition, advancements in AI detection tools must keep pace with the rapidly evolving nature of disinformation tactics. A collaborative effort across sectors is essential to ensure that these measures are effective in maintaining the integrity of online information ecosystems and protecting against the harms of misinformation.

# 7 References


Abrar, A., Abdul, W., & Ghouzali, S. (2021). Secure Image Authentication Using Watermarking and Blockchain. Intelligent Automation & Soft Computing, 28(2), 577–591. https://doi.org/10.32604/iasc.2021.016382

Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S.-N. (2020). Detecting Deep-Fake Videos From Appearance and Behavior. https://doi.org/10.1109/wifs49906.2020.9360904

Aghajari, Z., Baumer, E. P. S., & DiFranzo, D. (2023). Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–34. https://doi.org/10.1145/3579520

Ahmad, A. R., & Murad, H. R. (2020). The Impact of Social Media on Panic During the COVID-19 Pandemic in Iraqi Kurdistan: Online Questionnaire Study. Journal of Medical Internet Research, 22(5), e19556. https://doi.org/10.2196/19556

Ahmed, & Rasul. (2023). Examining the association between social media fatigue, cogni-

tive ability, narcissism and misinformation sharing: Cross-national evidence from eight countries | Scientific Reports. https://www.nature.com/articles/s41598-023-42614-z

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. https://doi.org/10.1007/s13278-023-01028-5

Alaphilippe, A., Gizikis, A., Hanot, C., & Bontcheva, K. (2019). Automated tackling of disinformation. EU DisinfoLab. https://www.disinfo.eu/publications/automated-tackling-of-disinformation

Al-Hawamleh, A. M. (2024). Investigating the multifaceted dynamics of cybersecurity practices and their impact on the quality of e-government services: Evidence from the KSA. Digital Policy, Regulation and Governance, 26(3), 317–336. https://doi.org/10.1108/DPRG-11-2023-0168

Amazeen, M. A., Krishna, A., & Eschmann, R. (2022). Cutting the Bunk: Comparing the Solo and Aggregate Effects of Prebunking and Debunking Covid-19 Vaccine Misinformation. Science Communication, 44(4), 387–417. https://doi.org/10.1177/10755470221111558

Arambul, N., Sraboni, S., Chukwunweike, J., & Olagoke, A. (2023). Exploring the Association between Trust in Healthcare Entities and Exposure to Emerging Health Misinformation in Nebraska: A Pilot Study. Human Factors and Ergonomics Society, 67(1), 1731–1734.

Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-033

Badari, A., & Chaudhury, A. (2021). An Overview of Bitcoin and Ethereum White-Papers, Forks, and Prices. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3841827

Badrinathan, S. (2021). Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. American Political Science Review, 115(4), 1325–1341. https://doi.org/10.1017/S0003055421000459

Banner, N. (2022). NHS data breaches: A further erosion of trust. Bmj, 377.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right. Psychological Science, 26(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Beaumont, R. (2022). Clip retrieval system. https://rom1504.github.io/clip-retrieval

Benet, J. (2014). IPFS - Content Addressed, Versioned, P2P File System. https://doi.org/10.48550/arxiv.1407.3561

Borges do Nascimento, I. J., Pizarro, A. B., Almeida, J. M., Azzopardi-Muscat, N., Gonçalves, M. A., Björklund, M., & Novillo-Ortiz, D. (2022). Infodemics and health misinformation: A systematic review of reviews. Bulletin of the World Health Organization, 100(9), 544–561. https://doi.org/10.2471/BLT.21.287654

Boutilier, A. (2023, May 8). Liberals try to delay fight over privacy rules for political parties—National | Globalnews.ca. Global News. https://globalnews.ca/news/9681510/liberals-fight-privacy-rules/

Bronstein, M. V., & Vinogradov, S. (2021). Education alone is insufficient to combat online medical misinformation. EMBO Reports, 22(3), e52282. https://doi.org/10.15252/embr.202052282

Bruns, A. (2017). Echo chamber? What echo chamber? Reviewing the evidence. QUT Eprints. https://eprints.qut.edu.au/113937

Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news. PLoS ONE, 16(6). https://doi.org/10.1371/journal.pone.0253717

Butterick, P. (2021). The Effects of NX-AS-401 on Methicillin Resistant Staphylococcus Aureus. https://doi.org/10.23889/suthesis.59037

Canadian Centre for Cyber Security, C. S. E. (2024, May). How to identify misinformation, disinformation, and malinformation. Canadian Centre for Cyber Security. https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. Psychological Science, 28(11), 1531–1546.

Chaniago, N., Sukarno, P., & Wardana, A. A. (2021). Electronic Document Authenticity Verification of Diploma and Transcript Using Smart Contract on Ethereum Blockchain. Register Jurnal Ilmiah Teknologi Sistem Informasi, 7(2), 149. https://doi.org/10.26594/register.v7i2.1959

Cheng, L.-C., Lu, W.-T., & Yeo, B. (2023). Predicting abnormal trading behavior from internet rumor propagation: A machine learning approach. Financial Innovation, 9(1), 3. https://doi.org/10.1186/s40854-022-00423-9

Chesney, R., & Citron, D. K. (2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (SSRN Scholarly Paper 3213954). https://doi.org/10.2139/ssrn.3213954

Chuai, Y., & Zhao, J. (2022). Anger can make fake news viral online. Frontiers in Physics, 10. https://doi.org/10.3389/fphy.2022.970174

Cialdini, R. (1994). Interpersonal influence. Persuasion: Psychological Insights and Perspectives, 195–217.

Cialdini, R. (2007). Influence: The psychology of persuasion (1st ed.).

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The Echo Chamber Effect on Social Media. Proceedings of the National Academy of Sciences, 118(9). https://doi.org/10.1073/pnas.2023301118

Cresci, S. (2020). A Decade of Social Bot Detection. Communications of the Acm, 63(10), 72–83. https://doi.org/10.1145/3409116

Cresci, S., Pietro, R. D., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The Paradigm-Shift of Social Spambots. https://doi.org/10.1145/3041021.3055135

Dai, H.-N., Zheng, Z., & Zhang, Y. (2019). Blockchain for Internet of Things: A Survey. Ieee Internet of Things Journal, 6(5), 8076–8094. https://doi.org/10.1109/jiot.2019.2920987

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1166120

Deb, A., Luceri, L., Badaway, A., & Ferrara, E. (2019). Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. https://doi.org/10.1145/3308560.3316486

Delhi, S. F. in N. (2019, April 1). Facebook bans

hundreds of pages in run-up to Indian elections. Financial Times (FT.Com). http://global.factiva.com/redir/default.aspx?P=sa&an=FTCOM00020190401ef41006pp&cat=a&ep=ASE

DeVerna. (2024). Identifying and characterizing superspreaders of low-credibility content on Twitter | PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0302201

Di Domenico, G., & Ding, Y. (2023). Between brand attacks and broader narratives: How direct and indirect misinformation erode consumer trust. Current Opinion in Psychology, 54, 101716. https://doi.org/10.1016/j.copsyc.2023.101716

Di Meco, L., & Wilfore, K. (2021, March 8). Gendered disinformation is a national security problem. Brookings. https://www.brookings.edu/articles/gendered-disinformation-is-a-national-security-problem/

Diaz Ruiz, C. (2023). Disinformation on digital media platforms: A market-shaping approach. New Media & Society, 14614448231207644. https://doi.org/10.1177/14614448231207644

Dou, Y., Liu, Z., Li, S., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing Graph Neural Network-Based Fraud Detectors Against Camouflaged Fraudsters. https://doi.org/10.1145/3340531.3411903

Dubois, E., & Blank, G. (2018). The Echo Chamber Is Overstated: The Moderating Effect of Political Interest and Diverse Media. Information Communication & Society, 21(5), 729–745. https://doi.org/10.1080/1369118x.2018.1428656

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. Nature Communications, 14(1), 62. https://doi.org/10.1038/s41467-022-35576-9

Edwards, A. (2013). (How) do participants in online discussion forums create 'echo chambers'?: The inclusion and exclusion of dissenting voices in an online forum about climate change. J. Argumentation Context 2, 127–150.

European Commission. (2022, June 16). 2022 Strengthened Code of Practice on Disinformation | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

Fedeli, A. M., Jeffrey Gottfried, Galen Stocking, Mason Walker and Sophia. (2019, June 5). Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed. Pew Research Center. https://www.pewresearch.org/journalism/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

Feng, Q., He, D., Zeadally, S., Khan, M. K., & Kumar, N. (2019). A Survey on Privacy Protection in Blockchain System. Journal of Network and Computer Applications, 126, 45–58. https://doi.org/10.1016/j.jnca.2018.10.020

Ferrara, E. (2017). Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. First Monday. https://doi.org/10.5210/fm.v22i8.8005

Fong, B. (2021). Analysing the behavioural finance impact of "fake news" phenomena on financial markets: A representative agent model and empirical validation. Financial Innovation, 7(1), 53. https://doi.org/10.1186/

s40854-021-00271-z

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the Corrective Effects of Media-Based Political Fact Checking? The Role of Contextual Cues and Naïve Theory. Journal of Communication, 63(4), 617–637. https://doi.org/10.1111/jcom.12038

Georgacopoulos, C., & Mores, G. (2020, July). How Fake News Affected the 2016 Presidential Election. LSU Faculty Websites. https://faculty.lsu.edu/fakenews/elections/sixteen.php

Gilbert, E., Bergstrom, T., & Karahalios, K. (2009). Blogs are echo chambers: Blogs are echo chambers in 42nd Hawaii International Conference on System Sciences. IEEE Computer Society, 1–10.

Global Knowledge. (2024). Cybersecurity Glossary of Terms. http://www.globalknowledge.com/ca-en/topics/cybersecurity/glossary-of-terms/

Goldsmith, Benjamin E. & Horiuchi, Yusaku. (2023). Does Russian election interference damage support for US alliances? The case of Japan -. https://journals.sagepub.com/doi/10.1177/13540661221143214

Goldstein, J. A., Chao, J., & Grossman, S. (2024). How persuasive is ai-generated propaganda? PNAS Nexus, 3(2), 34.

Golosova, J., & Romānovs, A. (2018). The Advantages and Disadvantages of the Blockchain Technology. https://doi.org/10.1109/aieee.2018.8592253

Gorwa, R., & Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. Policy & Internet, 12(2), 225–248.

Graves, L., Nyhan, B., & Reifler, J. (2016). Understanding Innovations in Journalistic Practice: A Field Experiment Examining Motivations for Fact-Checking. Journal of Communication, 66(1), 102–138. https://doi.org/10.1111/jcom.12198

Grimme, C., Assenmacher, D., & Clever, L. (2018). Changing Perspectives: Is It Sufficient to Detect Social Bots? 445–461. https://doi.org/10.1007/978-3-319-91521-0_32

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

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

Grömping, M. (2014). 'Echo chambers': Partisan Facebook groups during the 2014 Thai election. Asia Pac. Media Educat., 24.

Guess, A., & Lyons, B. (2020). Misinformation, Disinformation, and Online Propaganda. In Social Media and Democracy: The State of the Field, Prospects for Reform. Cambridge University Press.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. Proceedings of the National Academy of Sciences, 117(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Guru, Prof. S. (2024). Blockchain Based Decentralized Storage System. International Journal

for Research in Applied Science and Engineering Technology, 12(4), 75–83. https://doi.org/10.22214/ijraset.2024.59674

Haliassos, A., Vougioukas, K., Petridis, S., & Pantić, M. (2021). Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection. https://doi.org/10.1109/cvpr46437.2021.00500

Harjani, T., Roozenbeek, J., Biddlestone, M., Van Der Linden, S., Stuart, A., Iwahara, M., Piri, B., Xu, R., Goldberg, B., & Graham, M. (2022). A practical guide to prebunking misinformation.

Harrison, J. (2024, May 31). LibGuides: Fake News: How Fake News Spreads. University of Victoria Libraries. https://libguides.uvic.ca/fakenews/how-it-spreads

He, B., Ahamad, M., & Kumar, S. (2023). Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation. Proceedings of the ACM Web Conference 2023, 2698–2709. https://doi.org/10.1145/3543507.3583388

Heo, G., Yang, D., Doh, I., & Chae, K. (2020). Design of Blockchain System for Protection of Personal Information in Digital Content Trading Environment. https://doi.org/10.1109/icoin48656.2020.9016501

Heo, G., Yang, D., Doh, I., & Chae, K. (2021). Efficient and Secure Blockchain System for Digital Content Trading. Ieee Access, 9, 77438–77450. https://doi.org/10.1109/access.2021.3082215

Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011). Transforming Auto-Encoders. 44–51. https://doi.org/10.1007/978-3-642-21735-7_6

Hoffman, M. R., Ibáñez, L., & Simperl, E. (2020). Toward a Formal Scholarly Understanding of Blockchain-Mediated Decentralization: A Systematic Review and a Framework. Frontiers in Blockchain, 3. https://doi.org/10.3389/fbloc.2020.00035

Horton, C. (2018, August 21). The simple but ingenious system Taiwan uses to crowdsource its laws. MIT Technology Review. https://www.technologyreview.com/2018/08/21/240284/the-simple-but-ingenious-system-taiwan-uses-to-crowdsource-its-laws/

Huang, X., & Yi, W. (2023). An Image Copyright Authentication Model Based on Blockchain and Digital Watermarking. https://doi.org/10.21203/rs.3.rs-3078286/v1

Hui, P.-M., Shao, C., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). The Hoaxy Misinformation and Fact-Checking Diffusion Network. Proceedings of the International Aaai Conference on Web and Social Media, 12(1). https://doi.org/10.1609/icwsm.v12i1.14986

Indapwar, A. (2020). E-Voting System Using Blockchain Technology. International Journal of Advanced Trends in Computer Science and Engineering, 9(3), 2775–2779. https://doi.org/10.30534/ijatcse/2020/45932020

Inter-Parliamentary Union. (2016). Sexism, harassment and violence against women parliamentarians [Dataset]. https://doi.org/10.1163/2210-7975_HRD-1021-2016006

Ireton, C., & Posetti, J. (2018). Journalism, fake news & disinformation: Handbook for journalism education and training. https://unesdoc.unesco.org/ark:/48223/pf0000265552

Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A.-H. M., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Chowdhury, K. I. A., Anwar, K. S., Chughtai, A. A., & Seale, H. (2020). COVID-19–

Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. https://doi.org/10.4269/ajtmh.20-0812

IT Governance. (2024). Cyber Security. IT Governance. https://itgovernance.co.uk/what-is-cybersecurity

ITU. (2024). Cybersecurity. ITU. https://www.itu.int:443/en/ITU-T/studygroups/com17/Pages/cybersecurity.aspx

Jones, C. M., Diethei, D., Schöning, J., Shrestha, R., Jahnel, T., & Schüz, B. (2023). Impact of Social Reference Cues on Misinformation Sharing on Social Media: Series of Experimental Studies. Journal of Medical Internet Research, 25(1), e45583. https://doi.org/10.2196/45583

Joseph, A. M., Fernandez, V., Kritzman, S., Eaddy, I., Cook, O. M., Lambros, S., Silva, C. E. J., Arguelles, D., Abraham, C., Dorgham, N., Gilbert, Z. A., Chacko, L., Hirpara, R. J., Mayi, B. S., & Jacobs, R. J. (2022). COVID-19 Misinformation on Social Media: A Scoping Review. Cureus, 14(4). https://doi.org/10.7759/cureus.24601

Kanan, T., Obaidat, A. T., & Al-Lahham, M. (2019). SmartCert BlockChain Imperative for Educational Certificates. https://doi.org/10.1109/jeeit.2019.8717505

Khoo, B., Phan, R. C. ?W., & Lim, C. H. (2021). Deepfake Attribution: On the Source Identification of Artificially Generated Images. Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, 12(3). https://doi.org/10.1002/widm.1438

Koetke, J., Schumann, K., Porter, T., & Smilo-Morgan, I. (2023). Fallibility Salience Increases Intellectual Humility: Implications for People's Willingness to Investigate Political Misinformation. Personality and Social Psychology Bulletin, 49(5), 806–820. https://doi.org/10.1177/01461672221080979

Konstantinou, L., Caraban, A., & Karapanos, E. (2019). Combating Misinformation Through Nudging. In D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, & P. Zaphiris (Eds.), Human-Computer Interaction – INTERACT 2019 (pp. 630–634). Springer International Publishing. https://doi.org/10.1007/978-3-030-29390-1_51

Korshunova, I., Shi, W., Dambre, J., & Theis, L. (2017). Fast Face-Swap Using Convolutional Neural Networks. https://doi.org/10.1109/iccv.2017.397

Kotras, B. (2020). Mass personalization: Predictive marketing algorithms and the reshaping of consumer knowledge. Big Data & Society, 7(2), 2053951720951581.

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. Psychological Science in the Public Interest, 21(3), 103–156. https://doi.org/10.1177/1529100620946707

Kshetri, N. (2018). 1 Blockchain's Roles in Meeting Key Supply Chain Management Objectives. International Journal of Information Management, 39, 80–89. https://doi.org/10.1016/j.ijinfomgt.2017.12.005

Lee, S. K., Sun, J., Jang, S., & Connelly, S. (2022). Misinformation of COVID-19 vaccines and vaccine hesitancy. Scientific Reports, 12(1), 13681. https://doi.org/10.1038/s41598-022-17430-6

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychological Science in the Public Interest, 13(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., & Van Der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. European Review of Social Psychology, 32(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-Ray for More General Face Forgery Detection. https://doi.org/10.1109/cvpr42600.2020.00505

Lu, Y.-J., & Li, C.-T. (2020). GCAN: Graph-Aware Co-Attention Networks for Explainable Fake News Detection on Social Media. https://doi.org/10.48550/arxiv.2004.11648

Matern, F., Rieß, C., & Stamminger, M. (2019). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. https://doi.org/10.1109/wacvw.2019.00020

Micallef, N., He, B., Kumar, S., Ahamad, M., & Memon, N. (2020). The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. 2020 IEEE International Conference on Big Data (Big Data), 748–757. https://doi.org/10.1109/BigData50022.2020.9377956

Milner, H. (2002). Civic literacy: How informed citizens make democracy work. UPNE.

Moore, K. N. (2016). The Use of Recollection Rejection in the Misinformation Paradigm. Applied Cognitive Psychology, 30(6), 992–1004. https://doi.org/10.1002/acp.3291

Moretto, M., Ortellado, P., Kessler, G., Vommaro, G., Rodriguez-Raga, J. C., Luna, J. P., Heinen, E., Cely, L. F., & Toro, S. (2022). People are more engaged on Facebook as they get older, especially in politics: Evidence from users in 46 countries. Journal of Quantitative Description: Digital Media, 2. https://doi.org/10.51685/jqd.2022.018

Morgan, C. A., Southwick, S. M., Steffian, G., Hazlett, G., & Loftus, E. F. (2013). Misinformation Can Influence Memory for Recently Experienced, Highly Stressful Events. International Journal of Law and Psychiatry, 36(1), 11–17. https://doi.org/10.1016/j.ijlp.2012.11.002

Mosleh, M., & Rand, D. (2021). Measuring exposure to misinformation from political elites on Twitter. OSF. https://doi.org/10.31234/osf.io/ye3pf

Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: A review of research in social media. International Journal of Data Science and Analytics, 13(4), 271–285. https://doi.org/10.1007/s41060-022-00311-6

Nadimpalli, A. V., & Rattani, A. (2022). GBDF: Gender Balanced DeepFake Dataset Towards Fair DeepFake Detection. https://doi.org/10.48550/arxiv.2207.10246

Nakamoto, N. (2017). Centralised Bitcoin: A Secure and High Performance Electronic Cash System. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3065723

National Institute of Standards and Technology. (2018). Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1 (NIST CSWP 04162018; p. NIST CSWP 04162018). National Institute of Standards and Technology. http://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey. https://doi.org/10.48550/arxiv.1909.11573

Nyhan, B., & Reifler, J. (2015). Does Correcting Myths About the Flu Vaccine Work? An Experimental Evaluation of the Effects of Corrective Information. Vaccine, 33(3), 459–464. https://doi.org/10.1016/j.vaccine.2014.11.017

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The Hazards of Correcting Myths About Health Care Reform. Medical Care, 51(2), 127–132. https://doi.org/10.1097/mlr.0b013e318279486b

OECD. (2022a). Building Trust and Reinforcing Democracy: Preparing the Ground for Government Action. OECD. https://doi.org/10.1787/76972a4a-en

OECD. (2022b). Good practice principles for public communication responses to mis- and disinformation (OECD Public Governance Policy Papers 30; OECD Public Governance Policy Papers, Vol. 30). https://doi.org/10.1787/6d141b44-en

Office of the Privacy Commissioner of Canada. (2013, April 4). Survey of Canadians on Privacy-Related Issues. https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2013/por_2013_01/

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises. Mis Quarterly, 37(2), 407–426. https://doi.org/10.25300/misq/2013/37.2.05

Online Safety Act (2023). https://bills.parliament.uk/bills/3137

Ortiz, S. M. (2020). Trolling as a Collective Form of Harassment: An Inductive Study of How Online Users Understand Trolling. Social Media + Society, 6(2), 205630512092851. https://doi.org/10.1177/2056305120928512

Ozair, M. (2023, November 22). Misinformation in the Age of Artificial Intelligence and What it Means for the Markets. Nasdaq. https://www.nasdaq.com/articles/misinformation-in-the-age-of-artificial-intelligence-and-what-it-means-for-the-markets

Pantazi, M., Papaioannou, K., & van Prooijen, J.-W. (2022). Power to the People: The Hidden Link Between Support for Direct Democracy and Belief in Conspiracy Theories. Political Psychology, 43(3), 529–548. https://doi.org/10.1111/pops.12779

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior Exposure Increases Perceived Accuracy of Fake News. Journal of Experimental Psychology, 147(12), 1865–1880. https://doi.org/: http://dx.doi.org/10.1037/xge0000465.supp

Piccolo, L. S. G., Joshi, S., Karapanos, E., & Farrell, T. (2019). Challenging Misinformation: Exploring Limits and Approaches. In D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, & P. Zaphiris (Eds.), Human-Computer Interaction – INTERACT 2019 (pp. 713–718). Springer International Publishing. https://doi.org/10.1007/978-3-030-29390-1_68

Pielemeier, J. (2020). Disentangling Disinformation: What Makes Regulating Disinformation So Difficult? https://doi.org/10.26054/0D-CJBV-FTGJ

Pilditch, T. D., Roozenbeek, J., Madsen, J. K., & Van Der Linden, S. (2022). Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. Royal Society Open Science, 9(8), 211953. https://doi.org/10.1098/rsos.211953

Pocol, A., Istead, L., Siu, S., Mokhtari, S., & Kodeiri, S. (2024). Seeing is No Longer Believing: A Survey on the State of Deepfakes, AI-Generated Humans, and Other Nonveridical Media.

In B. Sheng, L. Bi, J. Kim, N. Magnenat-Thalmann, & D. Thalmann (Eds.), Advances in Computer Graphics (pp. 427–440). Springer Nature Switzerland.

Poredi, N., Chen, Y., Li, X., & Blasch, E. (2023). Enhance Public Safety Surveillance in Smart Cities by Fusing Optical and Thermal Cameras. https://doi.org/10.23919/fusion52260.2023.10224117

Poredi, N., Sudarsan, M., Solomon, E., Nagothu, D., & Chen, Y. (2024). Generative Adversarial Networks-Based AI-generated Imagery Authentication Using Frequency Domain Analysis. https://doi.org/10.1117/12.3013240

Portmann, E. (2018). Rezension „Blockchain: Blueprint for A New Economy“. HMD Praxis Der Wirtschaftsinformatik, 55(6), 1362–1364. https://doi.org/10.1365/s40702-018-00468-4

Posetti, J., & Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation.

Post, K. (2024, August 1). Ukraine and the Frontlines of the War on Disinformation—Foreign Policy Research Institute. Foreign Policy Research Institute. https://www.fpri.org/article/2024/08/ukraine-and-the-frontlines-of-the-war-on-disinformation/

Pretus, C., Servin-Barthet, C., Harris, E. A., Brady, W. J., Vilarroya, O., & Van Bavel, J. J. (2023). The role of political devotion in sharing partisan misinformation and resistance to fact-checking. Journal of Experimental Psychology: General, 152(11), 3116–3134. https://doi.org/10.1037/xge0001436

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo Chambers on Facebook. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2795110

Rich, T. S., MILDEN, I., & Wagner, M. T. (2020). Research Note: Does the Public Support Fact-Checking Social Media? It Depends Who and How You Ask. https://doi.org/10.37016/mr-2020-46

Robertson, C. T., Mourão, R. R., & Thorson, E. (2020). Who Uses Fact-Checking Sites? The Impact of Demographics, Political Antecedents, and Media Use on Fact-Checking Site Awareness, Attitudes, and Behavior. The International Journal of Press/Politics, 25(2), 217–237. https://doi.org/10.1177/1940161219898055

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions. European Psychologist, 28(3), 189–205. https://doi.org/10.1027/1016-9040/a000492

Roozenbeek, J., & Van Der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. Palgrave Communications, 5(1), 65. https://doi.org/10.1057/s41599-019-0279-9

Santos-D'amorim, K., & Miranda, M. K. F. de O. (2021). Misinformation, Disinformation, and Malinformation: Clarifying the Definitions and Examples in Disinfodemic Times. Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência Da Informação, 26. https://www.redalyc.org/journal/147/14768130011/html/

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. Proceedings of the National Academy of Sciences - PNAS. https://doi.org/10.1073/pnas.1805871115

Shah, S. B., Surendrabikram, T., Acharya, A., Rauniyar, K., Poudel, S., Jain, S., Masood, A., & Naseem, U. (2024). Navigating the Web of Dis-

information and Misinformation: Large Language Models as Double-Edged Swords. IEEE Access. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10540581&tag=1

Sikder, O., Smith, R. E., Vivo, P., & Livan, G. (2020). A minimalistic model of bias, polarization and misinformation in social networks. Scientific Reports, 10(1), 5493.

SK, S., Tellabati, D. S. H., Yetukuri, S., Ram, H. S. S., & Sivalasetty, R. P. (2024). Decentralization of Webhosting Using Blockchain. International Journal for Research in Applied Science and Engineering Technology, 12(3), 1893–1900. https://doi.org/10.22214/ijraset.2024.59232

Solomon, E. (2023). Face anti-spoofing and deep learning based unsupervised image recognition systems.

Solomon, E., & Cios, K. J. (2023). Hdlhc: Hybrid face anti-spoofing method concatenating deep learning and hand-crafted features. 2023 IEEE 6th International Conference On Electronic Information And Communication Technology (ICEICT) IEEE, 470–474.

Solomon, E., Woubie, A., & Cios, K. J. (2022). UFace: An Unsupervised Deep Learning Face Verification System. Electronics, 11(23), 3909. https://doi.org/10.3390/electronics11233909

Statistics Canada. (2023). Concerns with misinformation online, 2023 [Dataset]. https://www150.statcan.gc.ca/n1/daily-quotidien/231220/dq231220b-eng.htm

Statistics Canada. (2024). Trust in media and main source of news by gender and province [Dataset]. [object Object]. https://doi.org/10.25318/4510010201-ENG

Su, Z., & Agyingi, E. (2024). Modeling the Impact of Misinformation on the Transmission Dynamics of COVID-19. AppliedMath, 4(2), Article 2. https://doi.org/10.3390/appliedmath4020029

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of Health Misinformation on Social Media: Systematic Review. Journal of Medical Internet Research, 23(1), e17187. https://doi.org/10.2196/17187

Sunstein, C. R. (1999). The Law of Group Polarization. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.199668

Thai, M. T., Wu, W., & Xiong, H. (2016). Big Data in Complex and Social Networks. CRC Press.

Vardhan, H., Rohan, L., Reddy, A. K., & Mourya, M. (2023). Mitigating Counterfeiting Using Blockchain Enabled Product Authentication. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11(IV), Available online.

Varlamis, I., Michail, D., Glykou, F., & Tsantilas, P. (2022). A Survey on the Use of Graph Convolutional Networks for Combating Fake News. Future Internet, 14(3), 70. https://doi.org/10.3390/fi14030070

Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread of True and False News Online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Vraga, E. K., Kim, S. C., Cook, J., & Bode, L. (2020). Testing the Effectiveness of Correction Placement and Type on Instagram. The International Journal of Press/Politics, 25(4), 632–652. https://doi.org/10.1177/1940161220919082

Waldman, A. E. (2018). The Marketplace of Fake News. 20. https://core.ac.uk/outputs/210555896/?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decora-

tion-v1

Walker, A. S. (2019). Preparing Students for the Fight Against False Information With Visual Verification and Open Source Reporting. Journalism & Mass Communication Educator, 74(2), 227–239. https://doi.org/10.1177/1077695819831098

Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. H. (2023). A survey on ChatGPT: AI-generated contents, challenges, and solutions. IEEE Open Journal of the Computer Society.

White, N. R., & Roberts, J. V. (1985). Criminal intent: The public's view. Canadian Journal of Criminology, 27(4), 455–465.

Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2020). Graph Neural Networks in Recommender Systems: A Survey. https://doi.org/10.48550/arxiv.2011.02260

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. 5907–5915.

Zheng, Z., Xie, S., Dai, H. N., Chen, X., & Wang, H. (2018). Blockchain Challenges and Opportunities: A Survey. International Journal of Web and Grid Services, 14(4), 352. https://doi.org/10.1504/ijwgs.2018.095647

Zheng, Z., Xie, S., Dai, H.-N., Chen, X., & Wang, H. (2017). An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends. https://doi.org/10.1109/bigdatacongress.2017.85

Zhou, K., Šćepanović, S., & Quercia, D. (2024). Characterizing Fake News Targeting Corporations. Proceedings of the International AAAI Conference on Web and Social Media, 18, 1818–1832. https://doi.org/10.1609/icwsm.v18i1.31428

Zhou, Y., & Shen, L. (2022). Confirmation bias and the persistence of misinformation on climate change. Communication Research, 49(4), 500–523.

Human-Centric
Cybersecurity
Partnership