

Examining Misinformation and Deep Fakes

Namosha Veerasamy, Zubeida Khan and Danielle Badenhorst

CSIR, Pretoria, South Africa

nveerasamy@csir.co.za

zdawood@csir.co.za

dbadenhorst@csir.co.za

Abstract: Misinformation in the form of deep fakes and phishing links can not only spread false information but can only be used as a weapon in the hands of cyber criminals. To combat this problem, the authors investigate fake news and misinformation, in a South African context. In the paper, the use of cyber scams that contain misinformation will also be unpacked. This aims to create an awareness and defensive approach to tackling emerging cyber threats that prey on misinformation. This paper tackles a growing concern by examining the pervasiveness of fake news by looking into the extent that fake news infiltrates various media channels and its potential impact on public perception and decision-making. The paper will also explore the anatomy of fake news by dissecting the common tactics and strategies employed by purveyors of fake news and highlight red flags that can help the public identify misinformation. Maintaining academic integrity is pivotal to the research and publication community. This paper will also promote the use of trusted sources and verification of information. The paper aims to promote media literacy by sharing strategies to enhance media literacy and critical thinking skills, empowering individuals to discern credible information from misleading content. This paper proposes a human-centric framework to empower individuals in South Africa to become discerning consumers of information. Recognizing the limitations of Artificial Intelligence (AI)-based detection methods and the unique challenges of the South African context (multilingualism, resource constraints), the framework emphasizes critical thinking and media literacy skills. It outlines a step-by-step process for evaluating information sources, including source credibility analysis, content verification, and cross-referencing. The effectiveness of the framework is demonstrated by a relevant use-case.

Keywords: Misinformation, Deep Fakes, Fake News, Disinformation, Mal-Information, Artificial Intelligence, Cybersecurity, Cybersecurity Learning Factory

1. Introduction

"Fake news is not a new phenomenon. It has been around since news became a concept 500 years ago with the invention of print—a lot longer, in fact, than verified, "objective" news, which emerged in force a little more than a century ago. From the start, fake news has tended to be sensationalist and extreme, designed to inflame passions and prejudices (Soll 2016)". Fake news can be extremely dangerous, powerful and persistent. Wild fake stories can promote prejudice, distrust and circulate unverified content. Thrilling counts of monsters, politicians and criminals have been provoking and enthralling audiences. Trust in news establishments can be questionable and may continue to decline. Fake news has now become a powerful force as we are bombarded with content without any accuracy and objectivity. As fake news grows, there is also associated terminology like misinformation, disinformation, mal-information. These concepts are examined next.

Misinformation refers to false or inaccurate information, especially that which is deliberately intended to deceive (Oxford, 2024).

Disinformation is a form of propaganda involving the dissemination of false information with the deliberate intent to deceive or mislead (Oxford b 2024).

Mal-information is information that is based on reality but is used to inflict harm on a person, organisation or country such as catphishing or swatting (Mediadefence.org 2024)

Fake news are false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke (Cambridge 2024)

Fake news is false information that has been purposefully crafted, contains sensationalism, can be emotionally charged, misleads or fabricates information that mimic mainstream news. Fake news can be wholly fabricated or a mix of fact and fiction.

According to Wardle, there are seven types of fake news (Wardle 2017):

- Satire Or Parody ("No Intention To Cause Harm But Has Potential To Fool")
- False Connection ("When Headlines, Visuals Or Captions Don't Support The Content")
- Misleading Content ("Misleading Use Of Information To Frame An Issue Or An Individual")
- False Context ("When Genuine Content Is Shared With False Contextual Information")

- Impostor Content ("When Genuine Sources Are Impersonated" With False, Made-Up Sources)
- Manipulated Content ("When Genuine Information Or Imagery Is Manipulated To Deceive", As With A "Doctored" Photo)
- Fabricated Content ("New Content Is 100% False, Designed To Deceive And Do Harm")

The proliferation of fake news and misinformation can have vast consequences. Some examples include (University of Exeter Library 2024):

- Distrust in the media
- Undermining the democratic process
- Platforms for harmful conspiracy theories and hate speech
- Spread of false or discredited science

Information today has grown in volume, variety, veracity, value and velocity. In today's information landscape, the spread of fake news poses significant threats to individuals, organisations, and society at large. The challenges posed by the proliferation of fake news in the digital age is on the rise. The use of misinformation and fake news can have dire consequences. Harmful misleading intentions, influence, creation of bias, manipulated content and propaganda are some of the negative outcomes. Various collaborative efforts can be applied in order to try and combat the issues of fake news and misinformation. This paper provides valuable insights on how to overcome this issue of fake news and misinformation in the growing digital era.

Fake news raises a serious concern about the continued reliability and truthfulness of information that form the underlying base for knowledge domains. With the rise of additional technologies like Artificial Intelligence, these platforms support the growth and development of deep fakes and other malicious tampering of information.

The remainder of the paper is structured as follows. The next section of the paper looks at existing works, followed by a section on the use of misinformation for the proliferation of cyber scams. Thereafter the paper delves into deep fakes and the use of AI in Section 4. The subsequent section, Section 5, explores the detection and battle against fake news. Finally, the authors conclude and discuss future work.

2. Detection of Misinformation at an Elementary Level

When fake news or misinformation is encountered, users may easily trust that the content is true, even though it may be inaccurate. Readers face an ongoing struggle to determine whether the content they find in digital channels is actually true reporting or a fabrication.

Diverse methodologies and frameworks to detect mis-, disinformation and fake news **can be applied**. This section provides a brief overview of common techniques for detection of misinformation and fake news.

Many frameworks attempt to calculate scores of credibility to distinguish real or fake news. Bhattarai et al. (2021) introduced an interpretable fake news detection framework based on the Tsetlin Machine, using conjunctive clauses to capture lexical and semantic properties of news text. This approach not only classifies news as fake or real, but also calculates a credibility score for fake news (Bhattarai et al., 2021). Zhang et al. (2019) proposed a text analytics-driven approach for fake news detection that transforms news from a document-based corpus into a topic and event-based representation. Fake topics and fake events can be detected with their two-layered approach, with a 92.49% classification accuracy and 94.16% recall (Zhang et al., 2019).

Other approaches consider tools that could be employed by laypeople or focus on cognitive aspects of fake news and its proliferation. Choy & Chong (2018) propose a LeSiE (Lexical Structure, Simplicity, Emotion) framework, with the goal of equipping laypeople with the ability to identify fake news without requiring complex tool. They consider the anomalous lexical structure, over-simplification and increased cognitive simplicity, as well as emotionally arousing material found in fake news artefacts. Batailler et al. (2021) discuss the application of signal detection theory to fake news, separating the ability to accurately distinguish real from fake news from biases in judging news veracity. Their framework considers the influence of partisan bias and cognitive reflection on fake news identification (Batailler et al., 2021).

The frameworks in the literature often rely on machine learning trained on specific datasets that may not reflect the multilingual and multicultural reality of South Africa. South Africa has a history of political and social tensions caused by misinformation. A human-centric approach could be more relevant and reachable for citizens to critically evaluate information.

A further abuse of misinformation is the generation of cyber scams. Trusting users may view adverts, phishing emails and other promotions without any knowledge that the content is fraudulent. The next section looks at the application of misinformation in the creation of cyber scams.

3. Cyber Scams Using Misinformation

Awareness also needs to be created on the use of misinformation and disinformation in the generation and proliferation of cyber scams. This section provides a brief overview of just a few of these scams to shed light on other malicious practices that benefit from misinformation and disinformation.

False information can also be used for false advertising and fraud. False information can be used to trick people into handing over money, credentials or other benefits. Scams make use of fraudulent and illegal means to confuse/deceive people. The next section looks at a few of these schemes that utilise false information and advertising as part of a dishonest scheme (Brooks 2023). Such scams deceptively pose as legitimate business practices.

Various phishing scams may contain misinformation in order to trick users into divulging their personal information. Phishing emails can claim to update user account details or verify certain credentials. Phishing emails may impersonate banks, service providers, online shopping sites and a host of organisations that users traditionally transact with.

Office supply scams occurs when scammers pose as legitimate companies offering office supplies. Fraudsters impersonate legitimate companies offering office supplies. Businesses place orders and pay, but the supplies never arrive.

Fake invoicing occurs when businesses are targeted with seemingly legitimate invoices and payment requests. Due to the high volume of invoices processed, companies can be vulnerable to these scams.

Charitable solicitations occur when charitable requests may stem from scammers that are trying to take advantage to steal money. The fundraising or sponsorship invite may be fake and form part of a fraudulent charity scam; this tends occur following a disaster or high-profile event.

With many of these scams, AI technology is the catalyst to create fake requests and deceive the users into believing that the invitation is legitimate. The next section looks at the use of AI in creating fake content.

4. AI, Deep Fakes and Fake News

AI plays a major role in the dissemination of fake news, mis- and disinformation. Potential offensive uses of AI in fake news includes the creation of deepfakes, and automated content generation (Cybenko & Cybenko, 2018). AI technologies, particularly deep learning algorithms, enable the creation of convincing deepfakes. Deepfakes are algorithmically produced, hyper-realistic fake images, videos or audio recordings. Deepfakes have various applications, ranging from entertainment and education to malicious uses such as spreading misinformation, propaganda, election rigging and more.

Deepfakes have been used to create fake videos of political leaders, potentially influencing voter behaviour and destabilising political environments. For example, fabricated videos of politicians saying or doing things they never did could sway public opinion (Vaccari & Chadwick, 2020). Deepfake technology allows for the creation of content that can be rapidly spread, undermining trust in media institutions, which can be particularly damaging when used to spread propaganda or false narratives to large audiences (Rana et al., 2022). Beyond politics, deepfakes pose threats to individuals' privacy and security, including potential harassment and blackmail (Rana et al., 2022; Pantserov, 2020).

Furthermore, AI can be utilised to aid automated content generation; fake news articles or social media posts mimicking the style or appearance of legitimate news can be automated, making it difficult to individuals to discern between the validity of the content (Cybenko & Cybenko, 2018). Social media bots and amplification can be AI-driven and leveraged to amplify fake news stories, providing undue credibility. These bots can like, share and spread misinformation across platforms at a scale impossible for humans, increasing the reach of fake news using social media engagement algorithms (Meyers et al., 2020). Further, AI algorithms can analyse vast amounts of data to identify biases or preferences of individual users, and then microtargeting these users with fake news stories that the user would be more likely to believe or share, further propagating misinformation (Kumar et al., 2021). Advanced NLP techniques allow AI to write or modify fake news stories, making them more compelling

or tailored to specific audiences. This can include altering headlines or content to make them more clickable or shareable, thereby increasing their spread (Mohseni et al., 2020).

AI tools can also help with the generation of malicious content. Users can phrase the request in an innocent manner but thereafter use the response to create various scams like phishing, fake invoicing, charitable solicitations, or office supply scams. AI tools now provide for a seamless approach to content creation for scams as the outputs are now phrased appropriately and remove the obstacles of poor grammar or other tell-tale signs.

5. Detecting Fake News

Combatting misinformation requires critical thinking skills, fact-checking, and responsible sharing of information. Since AI and deepfakes are used to assist in the generation of fake news, detecting fake news requires a multifaceted approach. While automated software can be helpful in identifying certain red flags, critical thinking remains essential. A skilled person should be able to analyze the source of the information, examining its credibility and potential biases. Ultimately, fostering a questioning mindset is the key solution in detecting fake news.

Therefore, the scope of this framework is about detecting fake news as a person, i.e., a human-centric approach, empowering individuals with the critical thinking skills and knowledge to discern credible information from misleading content. Human -centric approach focuses on the needs, preferences and experiences of human beings. When faced with material, content and information, users are engaged in a human-centric experience as they need to intuitively digest and make sense of the subject matter. Automated, programmable analysis is not always ideal as users need to rapidly look through the information and make judgements, decisions and responses.

Other studies use software-based detection approach. There are several unique characteristics about fake news in South Africa that deems a human-centric approach preferred:

- Fast-paced cyberthreat landscape: This shift in focus aims to equip users with the long-term ability to navigate the ever-evolving landscape of online information.
- Lack of resources: In a disadvantaged country such as South Africa, some individuals do not have access to high-speed internet, or advanced devices needed to run complex AI-based detection software.
- Multi-lingual environment: South Africa has eleven official languages. AI detection tools trained on one language might miss fake news spread in another. A human-centric approach equips South Africans to analyse information regardless of language.
- Local culture: Understanding the cultural nuances of South Africa can help in identifying fake news that might play on specific cultural anxieties or biases. For instance, South African communication styles can sometimes be indirect or use humour. Understanding these nuances can help South Africans avoid misinterpreting online content as fake news simply because it uses sarcasm or satire. A human-centric approach allows for this contextual understanding.
- Digital literacy: While most South Africans have mobile devices and intermittent internet access, they lack the technical skills to navigate complex online platforms or apps. For example, many users make use of online platforms like WhatsApp in South Africa as a means of rapid communication. However, users may still lack knowledge about whether the messages received can be trusted. Human-centric tools can get better outreach for South Africans in order to empower them to become more discretionary and not just simply spread messages without verifying it.

5.1 i-Fakt (Identify fake truth) Human-Centric Framework

The ability to discern fact from fiction has become a critical skill. The proposed human-centric framework I-Fakt empowers individuals to navigate the complexities of the digital age by fostering media literacy and critical thinking. The rationale behind selecting this approach is rooted in the unique socio-cultural and technological landscape of South Africa. This includes 11 official languages, and a digital divide. The factors for this are now provided.

- Socio-Cultural Context: South Africa's diverse linguistic and cultural environment calls for a flexible and adaptable framework that can be applied across different languages and cultural contexts, catering for unique cultural nuances faced by the citizens. AI-based tools, often trained on datasets from specific languages and cultural backgrounds, may not be effective for this.

- Resource and Technological Constraints: Many South Africans lack access to high-speed internet at all times, and advanced technological devices required for sophisticated AI-based tools and technologies. A human-centric approach, which uses readily available resources is more practical and sustainable in this context.
- Digital/ Media Literacy Levels: While mobile device usage is high, digital literacy levels vary significantly, across the country. A framework that prioritizes education by critical thinking and media literacy is better suited to improving the awareness around misinformation and disinformation.

By following the steps of the framework, one can develop the tools needed to become a discerning consumer of information. A high-level overview is provided in Figure 1, and each pillar is unpacked in the following section.

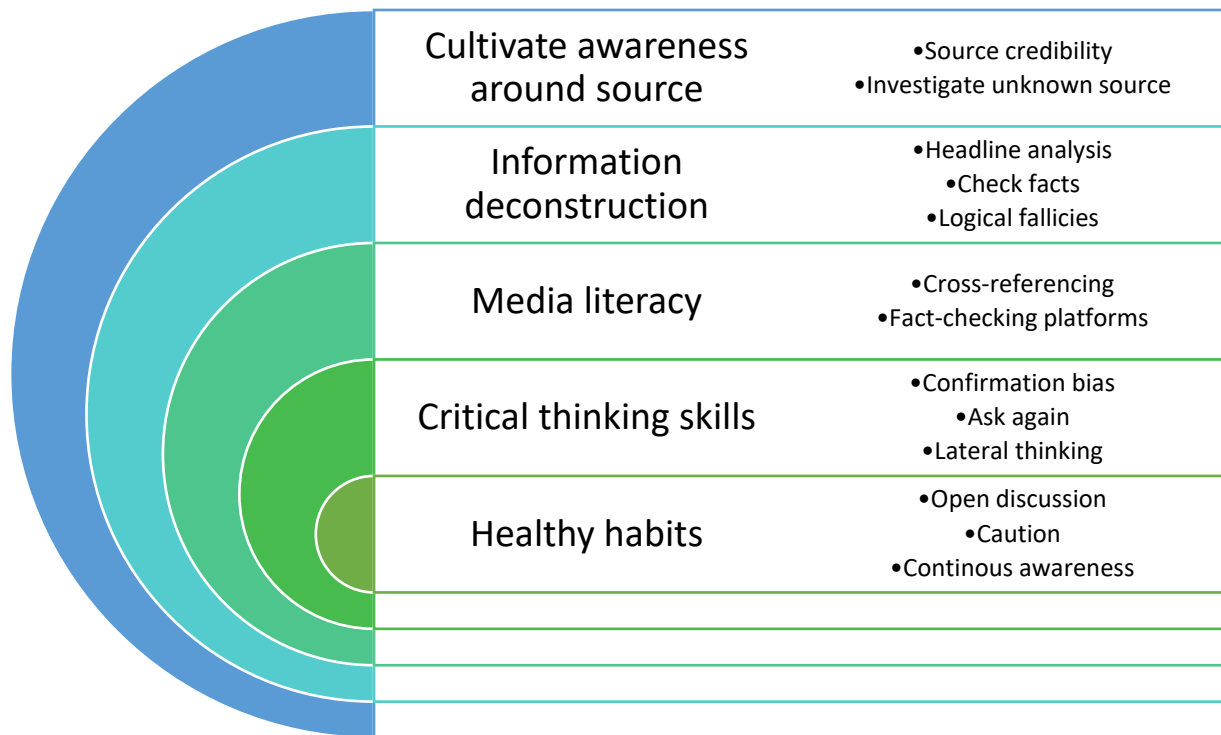


Figure 1: I-Fakt (Identify Fake Truth) Human-centric fake news detector framework

5.1.1 Cultivate Awareness Around Source

The onset of the framework lies in understanding where information and news stems from. The parameters of this pillar are as follows:

- Understanding Source Credibility: Learn to identify reputable news organizations with a history of fact-checking and ethical reporting. Explore resources that evaluate news source credibility. For instance,
- Investigate Unknown Sources: Before engaging with unfamiliar websites or social media pages, research their background and mission statement. Look for red flags like unusual domain extensions, unprofessional design, or lack of "About Us" or further information.

5.1.2 Information Destruction

Once the credibility of the source has been established, it is crucial to inspect the information itself, to understand the intent and veracity of the information.

- Headline analysis: Individuals need to be wary of sensationalized headlines with excessive exclamation points, strange grammar and uppercase letters. One must look for headlines that accurately reflect the content of the article.
- Check facts: Individuals should not solely rely on the information presented in the piece in question but should look for evidence to support claims, such as data from credible sources. For instance, consider a social media post claims that a new government policy will drastically reduce child grant payouts. This

sparks outrage and panic among many South Africans. An individual should check governmental policy websites to confirm this fact.

- Logical fallacies: Fake news commonly presents common logical fallacies such as appeals to emotion. For instance, a headline claiming that consuming South African herbs is a miracle cure for HIV/AIDS! Would appeal to individuals' sense of hope.

5.1.3 Media Literacy

This pillar of the framework empowers one to become a proactive participant in the digital information landscape by delving into techniques for verifying information beyond a single source.

- Cross-referencing: Individuals should not accept information at face value but should verify the information by searching for it from other established news outlets such as News24, The Daily Maverick, etc. If no reputable sources report the story, it could be fake.
- Fact-checking platforms: Utilize fact-checking websites like Snopes, PolitiFact, and Africa Check to verify information and debunk common fake news narratives.

5.1.4 Critical Thinking Skills

This pillar equips and individual to develop a questioning mind and foster intellectual curiosity.

- Confirmation bias: Individuals need to be aware of their own biases and how they might influence the interpretation of information. For instance, an individual, already having a negative perception of a political party, is more likely to accept negative information about the party in question at face value. One should seek out diverse perspectives to avoid falling victim to confirmation bias.
- Ask again: Approach online information sceptically. Individuals should question such as who is presenting such information, and do they have an agenda/ ulterior motive?
- Lateral thinking: Individuals should consider alternative explanations for the information presented. For instance, a certain celebrity did not attend the South African Music Awards. A voice note shared on messaging apps claims that he had been arrested for a serious crime. South Africans can apply lateral thinking by checking reputable news websites and social media accounts of the celebrity for any official statements before spreading fake news.

5.1.5 Healthy Habits

- Open discussion: Discuss this with peers and family members critically to consider whether the news is factual or not.
- Caution: Individuals should not share the piece of news without fact-checking. By sharing unverified information, individuals unwittingly contribute to the spread of fake news. This makes it more difficult for accurate information to reach people who need it.
- Continuous awareness: Regularly reading up and learning about current cybersecurity trends and engaging in discussions about media literacy and information verification reinforces healthy habits and strengthens the collective defence against misinformation. The cybersecurity learning factory [] could be rolled out around South African communities to encourage continuous awareness.

This framework empowers South Africans to become discerning consumers of information in the digital age. It emphasizes critical thinking and media literacy as key tools to assist in detecting fake news.

5.2 Use Case

A South African comes across the following post on X (formerly Twitter): *BREAKING! Disgraced again! [Politician's Name] caught in massive bribery scandal!!! Full story: [http://www.politiciansandal.co.za] #SouthAfrica #Corruption #PoliticalParty.*

As the citizen scrolls through Tiktok, he comes across a video, showing a clip with a voiceover: *"Did you hear about [Politician's Name]? Caught in a bribery scandal. See the full story at the link in my bio. #SouthAfrica #Corruption"*

An individual should consider each pillar of the framework and apply it to the post as follows.

- Cultivate Awareness Around Source: The source, on both X and Tiktok, leads to an unknown website with a sensational headline and lacks context. This raises red flags about credibility.
- Information Deconstruction: The headline is accusatory and uses exclamation points to evoke outrage. No evidence or details about the scandal are provided. The video on uses dramatic music and selective editing to amplify the claim without providing supporting evidence.
- Media Literacy: A search for the story on established news websites (in the South African context- News24, Daily Maverick, SABC News) reveals no mention of the incident. Checking Africa Check also shows no record of the claim in the above example.
- Critical Thinking Skills: Considering the political landscape, this accusation might be aimed at damaging the politician's reputation. Checking the politician's social media might reveal a response. One could also check if the politician has addressed the issue on their official TikTok or other social media accounts.
- Healthy Habits: Based on the analysis, the post and video is likely fake news. An individual should not share it and encourage others to be critical of online information.

The human-centric framework is an adaptable solution for combating misinformation. This use-case validates the framework's applicability across various contexts.

6. Conclusion

Knowledge on the topic is critical to preventing the spread of fake news and misinformation. Many readers may get caught up in the virality of certain feeds, posts and articles without any awareness that the information is sensationalism. The proliferation of fake news and misinformation poses a significant threat to South Africans. This research has examined the pervasiveness of it, its potential consequences, and strategies for detection. The authors found that AI plays a growing role in creating and disseminating fake news through techniques like deepfakes and automated content generation. Effectively detecting fake news as South Africans is challenging due to unique factors such as evolving threat landscape, cultural nuances, etc., and requires a human-centric approach based on critical thinking. This paper presented a human-centric framework empowering individuals to become discerning consumers of information in the digital age, alongside an illustrative use-case. Through this research, it is envisaged that users can be empowered with the techniques to fact check and verify information.

7. Future Work

Future work would involve extending the framework to include psychological factors. Understanding the psychological factors that make people susceptible to believing fake news can assist with improving judgment and reducing susceptibility to fake news and misinformation. Another direction is to consider the cybersecurity learning factory as a vehicle to demonstrate usage of the framework. The cybersecurity learning factory is an innovative educational platform designed to provide engaging and interactive learning experiences in cybersecurity. The aforementioned human-centric framework could be integrated into the Cybersecurity learning factory to include experiential learning into modules based on the framework. These modules can teach South Africans how to identify fake news sources, analyse information critically, and verify information through credible resources.

References

- Batailler, C., Brannon, S., Teas, P., & Gawronski, B., 2021. A Signal Detection Approach to Understanding the Identification of Fake News. *Perspectives on Psychological Science*, 17, pp. 78 - 98. <https://doi.org/10.1177/1745691620986135>.
- Bhattarai, B., Granmo, O., & Jiao, L. (2021). Explainable Tsetlin Machine Framework for Fake News Detection with Credibility Score Assessment. *ArXiv*, abs/2105.09114.
- Brooks C, 2023, Business Scams 101: Common Schemes and How to Avoid Them, Available at <https://www.business.com/security/business-scam-guide/>, Accessed 20240307
- Cambridge, 2024, Fake news, Available at <https://dictionary.cambridge.org/dictionary/english/fake-news>, Accessed 20240307)
- Choy, M., & Chong, M., 2018. Seeing Through Misinformation: A Framework for Identifying Fake Online News. *ArXiv*, abs/1804.03508.
- Cybenko, A., & Cybenko, G., 2018. AI and Fake News. *IEEE Intelligent Systems*, 33, pp. 1-5. <https://doi.org/10.1109/MIS.2018.2877280>.

- Gupta, A., Kumar, N., Prabhat, P., Gupta, R., Tanwar, S., Sharma, G., Bokoro, P., & Sharma, R., 2022. Combating Fake News: Stakeholder Interventions and Potential Solutions. *IEEE Access*, 10, pp. 78268-78289. <https://doi.org/10.1109/access.2022.3193670>.
- Mediadefence.org, 2024, Misinformation, Disinformation and Mal-information, Available at <https://www.mediadefence.org/ereader/publications/introductory-modules-on-digital-rights-and-freedom-of-expression-online/module-8-false-news-misinformation-and-propaganda/misinformation-disinformation-and-mal-information/#footnote--7>, Accessed 20230307.
- Murayama, T., Wakamiya, S., Aramaki, E., & Kobayashi, R., 2021. Modeling the spread of fake news on Twitter. *PLoS ONE*, 16. <https://doi.org/10.1371/journal.pone.0250419>.
- Oxford, 2024, Misinformation, Available at <https://www.google.com/search?client=firefox-b-e&q=definitions+for+misinformation>, Accessed 20240307
- Oxford b, 2024, Disinformation, Available at <https://www.google.com/search?client=firefox-b-e&q=disinformation+meaning>, Accessed 20240307
- Pantserov, K., 2020. The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability., pp. 37-55. https://doi.org/10.1007/978-3-030-35746-7_3.
- Paschen, J., 2019. Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management*. <https://doi.org/10.1108/JPB-12-2018-2179>.
- Rana, M., Nobil, M., Murali, B., & Sung, A., 2022. Deepfake Detection: A Systematic Literature Review. *IEEE Access*, 10, pp. 25494-25513. <https://doi.org/10.1109/access.2022.3154404>.
- Soll J, 18 December 2016, The Long and Brutual History of Fake News, *Politico Magazine*, Available at <https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/>, Accessed 20230307
- University of Exeter Library, 2024, Fake News, Available at <https://libguides.exeter.ac.uk/fakenews/consequences>, Accessed 20240307.
- Vaccari, C., & Chadwick, A., 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6. <https://doi.org/10.1177/2056305120903408>.
- Wardle C, 2017, Fake news. It complicated, Available at <https://firstdraftnews.org/articles/fake-news-complicated/>, Accessed 20230307.
- Zhang, C., Gupta, A., Kauten, C., Deokar, A., & Qin, X., 2019. Detecting fake news for reducing misinformation risks using analytics approaches. *Eur. J. Oper. Res.*, 279, pp. 1036-1052. <https://doi.org/10.1016/J.EJOR.2019.06.022>.