On Effective Ways to Warn People Against Misleading Contents*

Luigi Catuogno[1,*], Emanuel Di Nardo[1], Clemente Galdi[2] and Gennaro Ragucci[2]

[1]Universitá degli Studi di Napoli "Parthenope", Napoli (NA) - Italy
[2]Universitá degli Studi di Salerno, Fisciano (SA) - Italy

Abstract

Nowadays, means and services devoted to spreading news, information and opinions on the Internet are increasingly targeted by attacks aimed at inoculating malicious content with the aim of creating harmful effects on users, such as altering their perception of reality as well as influencing their opinions and choices. Plenty of promising solutions have been proposed to detect and classify such contents, according to their nature and scopes. Minor emphasis has been posed on choosing, for specific application contexts, the most effective way to warn users about dangerous contents. In this paper we first categorize different types of misleading information, identifying for of them peculiar aspects. Then, we report different methodologies that have been proposed to raise users' attention against potentially malicious contents. We argue that different alerting tools provide reasonable protection in specific contexts while no general solution is still available for guaranteeing users' security.

Keywords

Information Disorder, Social Engineering, User Warning

1. Introduction

Recent advances in the ICT have made available a variety of services and tools that enable people to communicate with a speed and quality never seen before. In addition, new media such as instant messaging, file sharing and social networks, have allowed them to access, exchange and spread information well beyond the domain of private communications. Nowadays, even individuals are able to reach a vast audience with self-produced contents aiming at promoting their creativity, beliefs, opinion or their own "version of facts".

On one hand, such level of connectivity exposes the users to criminals who try to convince them to release sensitive/financial information. On the other hand, we are witnessing an uncontrolled proliferation of information sources that challenge the prominence of traditional channels, without guaranteeing comparable quality and reliability. This phenomenon has raised numerous concerns as such an ungovernable and pervasive information flow might undermine the opinion formation process and the very reality perception among the public.

Regrettably, in recent years, the practice of deliberately disseminating misleading contents has been increasingly employed by malicious actors such as criminal organization or hostile states, to confuse and manipulate public opinion in order to pursuit political or economic objectives.

In this regard, the plague of *fake news*, which has raised the attention of institutions and academia in recent years (in particular since the COVID-19 pandemics), represents only the "tip of the iceberg" and is just one of the different types of manipulative campaigns based on the dissemination of malicious and misleading content. Other threats are related to the dissemination of propaganda with the aim of spreading hatred and mistrust in public opinion (e.g. *hate speeches*) or pushing the most vulnerable and marginalized individuals towards radical attitudes, that might motivate sensational and violent actions.

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy
*Corresponding author.
luigi.catuogno@uniparthenope.it (L. Catuogno); emanuel.dinardo@uniparthenope.it (E. Di Nardo); clgaldi@unisa.it (C. Galdi); gragucci@unisa.it (G. Ragucci)
0000-0002-6315-4221 (L. Catuogno); 0000-0002-6589-9323 (E. Di Nardo); 0000-0002-2988-700X (C. Galdi); 0009-0006-5265-5924 (G. Ragucci)

We observe that often the definitions of such contents (*i.e.*, *fake news*, *misinformation* and *disinformation*) are used interchangeably while, in reality, all these refer to well distinct criminal activities. Although they are framed within the more general context of *information disorder*, methods to recognize and thwart these activities have to cope with their wide variability in nature, form, mean and intents.

Many promising solutions have been proposed to this purpose. For example, detection of *hate speech* and radicalization messages mainly leverages opinion mining techniques and emotional analysis, whereas approaches that have proven to be effective for *fake news* recognition, range from tracking suspect contents back to their origin to analyzing their content seeking specific key words and sentences. Such solutions builds upon well-established technologies such as Deep Learning, Blockchains and Provenance analysis. Nevertheless, we argue that a comprehensive framework for mitigating the impact of misleading contents should necessary include mechanisms aimed at involving the public in the discernment process, with particular regard to those tools that should raise alerts or tag suspect information to trigger the users' attention. On this point, we notice that the current literature has not converged on a well-defined approach, yet, while many user warning systems are thought to fit specific context and applications. In addition, we highlight that within the industry (*e.g.*, among social media players) the adoption of user warning measures seems far from having reached any stability.

To the best of our knowledge, there is no comprehensive studies that consider all involved aspects at the same time.

2. Misleading contents classification

Controversial contents represent a complex and multidimensional phenomenon that profoundly affects communication processes and public perception of reality. They concern various forms and strategies, each characterized by specific methods of creation, dissemination, and impact. A comprehensive understanding of the terms associated with controversial contents is essential for analyzing the social, political, and ethical dynamics related to the circulation of information in digital and non-digital contexts.

The term information disorder is increasingly replacing "fake news" in discussions about information pollution. While fake news represents a specific type of information disorder, it is an inadequate term to capture the full complexity of the phenomenon. Even if the term means false or misleading information presented as news. It can be considered a generic term, in literature there is still no agreed definition of it and in 2017 Council of Europe Report (Executive Summary)[1], announced: they *refrein from using the term fake news for two reasons. First, it is woefully inadequate to describe the complex phenomena of information pollution. The term has also begun to be appropriated by politicians around the world to describe news organisations whose coverage they find disagreeable.* For these reasons Council of Europe Report (Part 1: Conceptual Framework)[1] proposes to use the following three terms that represent a breakdown of the different types of information disorder:

- **Mis-information** refers to inaccurate or incorrect information shared unintentionally without malicious intent. It is typically spread due to misunderstandings, misinterpretations or lack of verification, unlike disinformation. Instead other authors, e.g., [2], define misinformation as regardless of intent.

- **Dis-information** refers to false or misleading information that is intentionally created, disseminated, or promoted to cause harm, deceiving individuals or influencing public opinion for specific purposes. It is often strategically designed to manipulate perceptions, exacerbate divisions, or achieve political, economic, or ideological goals. Disinformation can take the form of fabricated narratives, or distorted interpretations of factual events, and it typically involves a deliberate element of deceit by its originators. Notice that have the authors in [2] defined misinformation regardless of intent, they consider disinformation as a subset of misinformation.

- **Mal-information** consists of genuine information that is shared with the intent to harm, manipulate, or mislead. Unlike disinformation, which involves falsehoods, malinformation is based on

verifiable truths but is used selectively or out of context to cause harm to individuals, groups, or institutions. Examples include making private information public (doxxing), leaks of confidential data, or the dissemination of truthful information in ways that provoke unnecessary outrage or fear.

So fake news can be categorized as either misinformation or disinformation, depending on the intent behind its creation. Propaganda is a related term used to describe information, true or false, that is spread to persuade an audience, often with a political motivation.

3. Attack techniques: Individual Compromise Versus Public Opinion Manipulation

Malicious contents have been used in the last decades in order to achieve several goals and for attacking different sets and types of targets. Traditional cyberattacks typically exploit device/software technical vulnerabilities in order to gain unauthorized (read and/or write) access to sensitive data. In principle, these types of attacks do require high-level technical skills on the attacker side and might not involve any human on the target side. Whenever malicious contents has to be conveyed to humans, as pointed out in [3] the attacker successful techniques leverages different social/psychological mechanisms and human vulnerabilities. However, depending on the type of attack target, the specific set of vulnerabilities and mechanisms lead to different types of attacks. We envision two major types of attack categories, the ones targeting a single or a small group of humans, the more classical social engineering, and the ones targeting public opinion, currently known as information disorder attacks.

3.1. Targeting individuals or small groups

Whenever the attack target is a single human being or a small group, typically the attack tries to convince the target user(s) to perform specific actions like releasing sensitive information or providing access to valuable resources. From the attacker point of view, these attacks are typically motivated by financial gain and the attacker is either a single person or a group of criminals.

The most widely known attack method in this context is *phishing*, where the attacker sends specifically crafted messages to users in order to induce them to execute the target operations. According to ENISA Threat Landscape 2024 [4], phishing is the still the preferred method for targeting users' credentials. Phishing attacks can be performed using different media like email, websites [5], SMS [6] (smishing), social media [7] or even phone calls (vishing). Messages used to manipulate victims' behaviour often create a sense of urgency, curiosity, or fear, compelling recipients to reveal credentials or click malicious links by overriding their critical thinking [8].

Although phishing is an old, well-established and known attack strategy, it has evolved over time into a number of more sophisticated variants. In its standard definition, phishing attacks consists in sending a message to users without any particular information on the specific user the message will reach. *Spear phishing* attacks are more targeted techniques since the messages that are sent to users are crafted using specific information on the user/set of users, generally obtained via OSINT techniques. The use of such information, e.g., published on the company website, allows the attacker to generete high credible messages. A further more specialized phishing variant is known as *whaling* that targets high value individuals like CEOs of CFOs. Clearly, the higher is the value of the target, the greater is the amount of work and information the attacker tries to gain in order to be successful in the attack. The main reason for which phishing attacks are still successful is attacker adaptability. Indeed, as public awareness increases, since new technologies are available, attacker crafts messages that use more complex psychological manipulations and adapting to currently available technology trends. For example, the 2025 APWG Phishing Activity Trends Report [9] highlights the increased use of QR codes embedded in phishing messages leading victims to phishing sites or malware downloads.

Generative AI is increasingly used for the [10, 11] the creation of tailored phishing campaigns, using information retrieved using OSINT techniques. These technologies tend to be a cheap, yet effective, solution for spear phishing attacks.

Another type of attack is known as *pretexting*, in which the attacker tries to create a a deceptive scenario on order to gain victim's trust in executing a specific operation. A typical example of pretexting may be a person who pretends to be a bank employee that needs to obtain access codes for home banking. There may be multiple interactions, using different communication means, between the victim and the attacker. Unlike phishing, that typically uses the urgency a psychological stress to indice the victim to perform the desired actions, pretexting tries to induce the victim to trust the attacker. A similar type of attack is *piggybacking* in which the attacker tries to gain the trust of the victim while entering a restricted (physical) area, e.g. pretending to be an employee whose badge is not working.

Baiting tries to induce the desired victim's behaviour by exploiting her curiosity promising free gifts like gift cards, free music/access to services and so on. These promises may induce the victim to follow a predefined path that typically leads to malware downloading or to provide sensitive information like credi cards data. Notice that baiting may also be implemented using physical objects, like digital devices intentionally left in public spaces. Along this line, a varian of baiting is *quishing* [12] that consists in leaving QR codes in public spaces with messages that leverage on victims curiosity. While baiting uses curiosity or desire, *scareware* use fear to indice a desired behaviour. A typical scareware attack consists in sending messages to a victim containing false alarms and proposing an easy way out to close the issue, e.g., a letter from the police threatening to arrest the victim if a fine is not paid on time using a link in the message.

3.2. Mass Manipulation

Information disorder techniques aims at shaping the public opinion perception or belief of a large (possibly targeted) audience, or of the general public. Example of information disorder campaigns are the one trying to influence political elections, polarizing the opinions in groups or discouraging participation in public events/campaigns.

In this field, disinformation campaigns play a central role. There are two key issues of such campaigns. The former is the generation of false information, either disinformation or malinformation in our original classification. The latter is the use of digital platforms and social media to spread such information, in order to achieve the original goal and pose significant threat to the public health, fair elections and so forth. These type of disinformation campaigns are typically much more effective than the traditional propaganda. The possibility of reaching a large number of subjects in very few time, of targeting specific groups of users with specifically crafted messages, the lack of borders and the possibility of using multiple (apparently different) sources leading toward the same thesis, along with the reduced capacity of people of distinguishing the truth from fabricated facts, make this tools very effective.

Computational propaganda [13] uses modern tools like algorithms, LLMs and big data to generate and distribute sequences of properly crafted messages with the intent of modify public opinion using social media as main communication mean. This type of propaganda is typically used by state actors or political parties to shape public opinion or discredit political opponents. One possible technique is astroturfing, the deliberate creation or manipulation of public opinion to give the appearance of grassroots support but is actually orchestrated and funded by a concealed entity. It applies to a particular agenda, organization, or product. This term is derived from AstroTurf, a brand of synthetic grass, symbolizing a manufactured facade of organic support. Astroturfing often employs fake online profiles, scripted testimonials, or coordinated campaigns to amplify specific viewpoints or suppress dissent. It is widely regarded as deceptive and unethical, as it undermines authentic public discourse and can distort decision-making processes in areas such as consumer behavior and politics [2]. Similarly, by *flooding* an attacker tries to spam social media to shape a narrative or drown out opposing viewpoints.

Deepfake [14] refers to synthetic media, typically videos or images, created using artificial intelligence to manipulate or fabricate content that appears real. Often generated through machine learning techniques such as generative adversarial networks (GANs), deepfakes can convincingly alter facial

expressions, voices, and other features. While they have legitimate applications in entertainment and education, deepfakes raise ethical and security concerns due to their potential use in spreading misinformation, blackmail, or political manipulation.

Fabricated content is a form of information disorder consisting of entirely false information deliberately created to mislead, deceive, or manipulate audiences. Unlike manipulate information (altered or distorted), fabricated content is entirely constructed without basis in fact or reality. It can manifest as fabricated news articles, fake social media posts, or fictional narratives masquerading as genuine reports, often intended to influence public opinion, incite emotions, or achieve political, financial, or social objectives.

Hate speech refers to communication, whether verbal, written, or symbolic, that denigrates, discriminates against, or incites hostility, violence, or prejudice toward individuals or groups based on characteristics such as race, religion, ethnicity, gender, sexual orientation, or disability. It is characterized by its intent or potential to harm societal cohesion and individual dignity and is often scrutinized within legal, ethical, and sociopolitical frameworks. Hate speech is distinct from other forms of controversial speech due to its capacity to fuel social tensions and perpetuate systemic inequality.

A *hoax* is a deliberate act of deception intended to mislead individuals or groups into believing something false. Hoaxes are premeditated and often designed to provoke emotional reactions, garner attention, or achieve financial or ideological gains. Hoaxes can take various forms, including fabricated news stories, fraudulent discoveries, or staged events, and are considered a significant source of misinformation.

3.3. Classifying mis/dis-information attacks

As usual, the classification of social engineering and information disorder attacks can be done using different categories. We have already partially mentioned the classification done in [3], where the authors classify social engineering attacks along three axis The first one is name *effect mechanisms*, that are persuasion, social influence, cognition & attitude & behavior, trust and deception, language & thought & decision, emotion and decision-making. The second feature are human vulnerabilities, i.e., cognition and knowledge, behavior and habit, emotions and feelings, human nature, personality traits, individual characters. The latter is the *attack method* that includes 16 attacks scenarios.

Since our focus is on alerting systems, we propose a classification whose primary axis is the cardinality of the target size, i.e, either one/small groups or public opinion. Indeed, the alerting systems need to act in completely different ways, e.g., blocking immediately a fraud directed to a specific user instead of marking as ”potentially fake” some post on a social media group.

Given the primary classification feature, we consider three more feature to be relevant for the purpose of defining the alerting systems. The first one is the objective of the attack. Example are financial gain, credential theft and so on. The primary attack objective on one hand provides a measure of the resources put in place to carry out the attack that, in turn, helps to identify the degree of specificity the attacker is using, e.g., phishing vs whaling. Furthermore, this dimension also provides an indicator on the urgency with which alerting/protection measures need to be raised. For example, users should be alerted immediately of a financial fraud, while the effect of fake news on group behaviour requires multiple events and more time.

Another important feature to be considered is the psychological factor that are used to induce users' behaviour. In general, attacks targeting individuals are more likely to use personal psychological factors, while attacks targeting the public opinion are more likely to use social influence and cognitive biases. The latter important feature to consider is clearly the attacks technique, that is crucial to identify the threat and alert the user.

4. Security warning systems

In Section 3, we introduced the different types of misleading content and highlighted the features that distinguish them on the basis of their target and goals. Such differentiation Such differentiation neces-

Characteristic	Individual-Targeted Social Engineering	Public Opinion-Targeted Social Engineering
Target	Single User or Small, Specific Group	Broad Audience or General Public
Attack Objective	Direct Financial Gain, Credential/-Data Theft, System Access, Malware Installation	Political/Societal Influence, Narrative Shaping, Belief/Behavioral Change, Discrediting Opponents
Psychological Factor	Urgency, Fear, Curiosity, False Trust, Greed, Shame	Cognitive Biases, Emotional Amplification, Social Proof, Authority, Consistency
Representative Techniques	Phishing (Spear, Whaling), Pretexting, Baiting, Scareware, Tailgating	Disinformation Campaigns, Computational Propaganda, Astroturfing, Deepfakes, Framing, Hate Speech, Hoax

sarily reflects in the corresponding warning systems. In particular, we observe that social engineering attacks show features that make their detection less error prone (*e.g.*, IP spoofing, phishing attempts). As a result, notification mechanisms can take more decisive actions, such as preventing the user to interact with the suspect content. In contrast, in information disorder campaigns malicious contents are spread through patterns that do not significantly deviate from the ones of legitimate contents (not, at least, as far as the final user can observe) on a small scale— thus requiring greater user involvement into identify and evaluating the trustworthiness of signaled contents.

4.1. Early measures against social engineering attacks

Since its advent, the use of the World Wide Web to provide services to a growing broad audience (often lacking of any technical expertise) and to support increasingly sensitive activities, such as healthcare, financial services, and citizenship-related services, has been accompanied by a significant rise in the intensity and severity of security threats. In particular, attack techniques based on *social engineering* have proven to be particularly challenging to be countered without involving the user in the process.

To this end, plenty of solutions aiming at helping users in recognizing threats from *e.g.* misleading pop-ups, phishing e-mails and counterfeit web pages were initially proposed. Such solutions relied on equipping e-mail agents and web browser with "security toolbars" that used to notify the users of potential risks. Furthermore, the rising adoption of SSL and HTTPS protocols provided indicators (*e.g.*, the state of the padlock icon) to alert users when they came across sites that could not be authenticated. The effectiveness of these warning systems has turned out to be quite limited[15, 16, 17].

In the following years, several field studies were carried out to examine how users respond to security warnings and to suggest ways to design more effective notification systems. Web browsers, which had become the main tool for accessing the internet, introduced a new generation of notification systems with improved performance. These active warning systems interrupt user navigation when a security risk is detected, requiring the user to take specific actions to address the issue. While these systems showed greater effectiveness, their performance varied widely depending on the type of user and the design and usability of their interface[18, 19].

4.2. Dealing with false reviews

Companies that sell products online often rely on customer review systems to promote their offerings. This exposes them to the risk of receiving fake reviews, that can jeopardize their image and their business. Here, we focus on strategies that can be suitable to handle suspect reviews once they have been detected, in order to mitigate their impact on new potential customers. For example, TripAdvisor removoves suspect contents and, possibly suspends the user who posted it [20, 21]. Other companies follow a more prudent and flexible strategy, by framing or marking controversial reviews and adopting progressive measures against potentially fraudulent users. In particular, Yelp [22] moves such content to a less evident position on the web site and do take them in account in computing scores and statistics.

However, within this application scenario, “censorship” has been criticized [23].

4.3. The contribution by social networks

Social networks, aware of their responsibility in the proliferation of controversial content, have implemented a range of measures over the years to combat this phenomenon.

Facebook (now Meta), was among the earliest players to recognize the impact of fake news. Initially, the platform relied on third-party fact-checking organizations to verify the accuracy of news reported by users. Subsequently, warning labels, as “Disputed” or “Rated false” tags, were introduced to mark potentially false or misleading content, as well as measures to reduce the visibility of such content in users’ feeds. However, the effectiveness of these measures has been the subject of debate[24]. Studies have shown that warning labels do not always deter users from sharing fake news. Moreover, by reducing the reach of some content, Facebook’s algorithms may still contribute to political polarization and the creation of “echo chambers” [25].

X (formerly Twitter) also adopted a similar approach and introduced warning labels for tweets containing false or misleading information. In addition, the platform implemented mechanisms to suspend those accounts that systematically spread disinformation. However, the provider has been criticized for the slowness and ineffectiveness of its measures. In particular, the platform’s algorithm has been accused of amplifying the spread of fake news and inciting hatred[26]. Afterwards, the platform chosen to move to a user reporting mechanisms called “community notes” [27].

YouTube has focused its efforts on removing content that violates its policies on disinformation, such as videos that promote conspiracy theories or false medical cures. In addition, YouTube has introduced information panels to provide context and verified information on controversial topics. Nevertheless, content moderation on YouTube remains challenging, due to the huge amount of videos uploaded daily. Furthermore, it has been shown that YouTube’s recommendation algorithm can contribute to the radicalization of users, suggesting videos carrying extremist or conspiracy messages [28].

TikTok, also focuses on removing content that violates its policies, such as videos inciting violence or discrimination. However, TikTok has faced criticism for its lack of transparency in moderation and for struggling to combat disinformation, especially during crises or elections [29].

5. Discussion

Warning systems essentially differ in how they trigger the users’ attention and prompt them to take actions; the content of the warning; and the subject of the notification.

Warning systems can issue *contextual* or *interstitial* notification. A contextual warning appears nearby the content it is referred to and does not restrict user activity. Marks, flags and frames fall in this category. Interstitial warnings interrupt the user and demands interaction. Such systems present dialog boxes or interactive elements that overlaps the underlying content, temporarily restricting access until a disclaimer has been viewed or a specific user action is taken.

Authors in [30] compares the performance of these two types of notification considered metrics include users’ interaction with the warning message (e.g. click rate, time spent in reading the disclaimer), additional users’ actions such as searching the web for alternative information and abandoning the reported contents. As expected, interstitial notifications promise better effectiveness, though their acceptation depends on their *friction*: a property that measure the perceived extent of the restrictions.

The content of notifications ranges from simple icons to labels such as ’not recommended’, or pop-up windows containing brief explanatory text and links to alternative sources of information related to the flagged content. These elements may be generated either by human moderation teams or automatically.

Warnings can be related to different aspects of the message, including: its content; its *emotional load*; the identity and the supposed reliability of the sender; or the trustworthiness of the source website [31].

In general, users appear to accept warning systems to prevent the spread of misleading contents as long as flagging criteria are transparent and explainable [32, 33]. Nevertheless, the effectiveness

of warning messages is strongly related to user demographic characteristics such as the age and the education level [34].

Misleading content appear to be majorly spread through social networks, due to their diffusion and pervasiveness. These platforms adopt different contrast strategies both on their own and relying on third party entities such as professional fact-checkers. In the latter case, these entities are allowed to access the data through specific APIs and according to certain policy. Both have not reached ad adequate stability yet. This makes difficult to implement warning mechanisms that achieve those adequate levels of usability and transparency that make them effective as expected. A possible solution to this problem, beyond regulatory improvements, may be the development of platform-independent warning system. Researches in this field are very rare though are worth of future investigation.

The design of a comprehensive framework for evaluating the performance of warning systes also looks an hot topic [35, 24]. Several works propose qualitative and quantitative evaluation criteria [30, 36]. However, we must notice that these proposals address different use cases and are still difficult to compare.

6. Conclusion

Currently available techniques for behavior manipulation can be essentially partitioned into two categories, based on the size of the set of users that are targeted by the attack. Attacks targeting single/small groups, tend to focus on immediate gain of sensitive/financial data. In contrast, whenever the intent is to alter social perspectives, attacks orchestrates widespread influence to shape collective narratives and achieve broader societal or political outcomes. The complexity of this landscape has been further increased by the introduction of tools that use artificial intelligence for content generation/manipulation.

Significant efforts have been devoted to the development of solutions for detecting misleading content and limiting its dissemination. In our opinion, however, much remains to be done in designing tools that actively involve users in this process.

Existing warning systems still have a limited effectiveness due several factors. In this paper we aim at highlight some of them. First, the development of any warning system strictly depends on the proprietary technologies of the different platforms (*e.g.*, social networks) and their policy about data dissemination. Second, the definition of comprehensive framework for effectiveness measurement is still due.

Acknowledgements

This work was supported by project IDA - Information Disorder Awareness (CUP D43C22003050001) - Spoke 2- SEcurity and RIghts in the CyberSpace - SERICS (PE00000014) - Under the NRRP MUR M4C2 I 1.3 program funded by the EU-NGEU; by project Privacy Preserving Data Analysis (PPDA)—Growing Resilient Inclusive and Sustainable—GRINS (PE00000018) Spoke 0 AND 2—under the NRRP MUR program funded by the EU—NGEU; and by a grant of the Università di Salerno.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, volume 27, Council of Europe Strasbourg, 2017.

[2] R. Di Pietro, S. Raponi, M. Caprolu, S. Cresci, Information Disorder, Springer International Publishing, Cham, 2021, pp. 7–64. URL: https://doi.org/10.1007/978-3-030-60618-3_2. doi:10.1007/978-3-030-60618-3_2.

[3] Z. Wang, H. Zhu, L. Sun, Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods, IEEE Access 9 (2021) 11895–11910. doi:10.1109/ACCESS.2021.3051633.

[4] European Union Agency for Cybersecurity (ENISA), Enisa threat landscape 2024, 2024. URL: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024, european Union Agency for Cybersecurity (EU body or agency).

[5] A. A. Ahmed, N. A. Abdullah, Real time detection of phishing websites, in: 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016, pp. 1–6. doi:10.1109/IEMCON.2016.7746247.

[6] Z. Ramzan, Phishing Attacks and Countermeasures, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 433–448. URL: https://doi.org/10.1007/978-3-642-04117-4_23. doi:10.1007/978-3-642-04117-4_23.

[7] T. N. Jagatic, N. A. Johnson, M. Jakobsson, F. Menczer, Social phishing, Communications of the ACM 50 (2007) 94–100.

[8] G. Desolda, L. S. Ferro, A. Marrella, T. Catarci, M. F. Costabile, Human factors in phishing attacks: A systematic literature review, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3469886. doi:10.1145/3469886.

[9] AntiPhishing Working Group, Inc., Phishing activity trends reports, 2024. URL: https://apwg.org/trendsreports/.

[10] M. Schmitt, I. Flechais, Digital deception: generative artificial intelligence in social engineering and phishing, Artificial Intelligence Review 57 (2024) 324. URL: https://doi.org/10.1007/s10462-024-10973-2. doi:10.1007/s10462-024-10973-2.

[11] A. M. Shibli, M. M. A. Pritom, M. Gupta, Abusegpt: Abuse of generative ai chatbots to create smishing campaigns, in: 2024 12th International Symposium on Digital Forensics and Security (ISDFS), 2024, pp. 1–6. doi:10.1109/ISDFS60797.2024.10527300.

[12] M. Rys, A. Ślusarek, K. Zieliński, Caught off guard: When experts fall for quishing, is awareness enough?, SECURITY AND PRIVACY 8 (2025) e486. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.486. doi:https://doi.org/10.1002/spy2.486. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/spy2.486.

[13] X. Cheng, L.-X. Yang, Q. Zhu, C. Gan, X. Yang, G. Li, Cost-effective hybrid control strategies for dynamical propaganda war game, IEEE Transactions on Information Forensics and Security 19 (2024) 9789–9802. doi:10.1109/TIFS.2024.3468903.

[14] A. Santha, S. L. Alarcon, C. Hochgraf, Deepfakes Generation Using LSTM Based Generative Adversarial Networks, Master's thesis, Rochester Institute of Technology, 2020.

[15] M. Wu, R. C. Miller, S. L. Garfinkel, Do security toolbars actually prevent phishing attacks?, in: Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006, pp. 601–610.

[16] S. E. Schechter, R. Dhamija, A. Ozment, I. Fischer, The emperor's new security indicators, in: 2007 IEEE Symposium on Security and Privacy (SP'07), IEEE, 2007, pp. 51–65.

[17] S. Egelman, L. F. Cranor, J. Hong, You've been warned: an empirical study of the effectiveness of web browser phishing warnings, CHI '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 1065–1074. URL: https://doi.org/10.1145/1357054.1357219. doi:10.1145/1357054.1357219.

[18] D. Akhawe, A. P. Felt, Alice in warningland: a large-scale field study of browser security warning effectiveness, in: Proceedings of the 22nd USENIX Conference on Security, SEC'13, USENIX Association, USA, 2013, p. 257–272.

[19] R. W. Reeder, A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, S. Egelman, An experience sampling study of user reactions to browser warnings in the field, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–13. URL: https://doi.org/10.1145/3173574.3174086. doi:10.1145/3173574.3174086.

[20] C. G. Harris, Decomposing tripadvisor: Detecting potentially fraudulent hotel reviews in the era of big data, in: 2018 IEEE international Conference on big knowledge (ICBK), IEEE, 2018, pp. 243–251.
[21] M. Mkono, 'troll alert!': Provocation and harassment in tourism and hospitality social media, Current Issues in Tourism 21 (2018) 791–804.
[22] D. Kamerer, Understanding the yelp review filter: An exploratory study, First Monday 19 (2014).
[23] U. M. Ananthakrishnan, B. Li, M. D. Smith, A tangled web: Should online review portals display fraudulent reviews?, Information Systems Research 31 (2020) 950–971.
[24] K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, M. Sandhu, R. Sang, R. Scholz-Bright, A. T. Welch, A. G. Wolff, A. Zhou, B. Nyhan, Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media, Political Behavior 42 (2020) 1073 – 1095. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061502088&doi=10.1007%2fs11109-019-09533-0&partnerID=40&md5=8454e4df888603d991fa2c56f142916a. doi:10.1007/s11109-019-09533-0, cited by: 327.
[25] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science 348 (2015) 1130–1132.
[26] K. Starbird, A. Arif, T. Wilson, Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations, Proceedings of the ACM on Human-Computer Interaction 3 (2019) 1–26.
[27] Y. Chuai, H. Tian, N. Pröllochs, G. Lenzini, Did the roll-out of community notes reduce engagement with misinformation on x/twitter?, Proc. ACM Hum.-Comput. Interact. 8 (2024). URL: https://doi.org/10.1145/3686967. doi:10.1145/3686967.
[28] M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, W. Meira Jr, Auditing radicalization pathways on youtube, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020) 131–141.
[29] C. E. Kirkpatrick, L. L. Lawrie, Tiktok as a source of health information and misinformation for young women in the united states: Survey study, JMIR Infodemiology 4 (2024) e54663. URL: https://infodemiology.jmir.org/2024/1/e54663. doi:10.2196/54663.
[30] B. Kaiser, J. Wei, E. Lucherini, K. Lee, J. N. Matias, J. Mayer, Adapting security warnings to counter online disinformation, in: 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 1163–1180.
[31] S. Lee, S. Afroz, H. Park, Z. J. Wang, O. Shaikh, V. Sehgal, A. Peshin, D. H. Chau, Explaining website reliability by visualizing hyperlink connectivity, in: 2022 IEEE Visualization and Visual Analytics (VIS), 2022, pp. 26–30. doi:10.1109/VIS54862.2022.00014.
[32] J. Kirchner, C. Reuter, Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness, Proceedings of the ACM on Human-computer Interaction 4 (2020) 1–27.
[33] Y.-L. Hsu, S.-C. Dai, A. Xiong, L.-W. Ku, Is explanation the cure? misinformation mitigation in the short term and long term, 2023. URL: https://arxiv.org/abs/2310.17711. arXiv:2310.17711.
[34] P. Mihai-Ionuț, E. Irina, Influence of the educational level on the spreading of fake news regarding the energy field in the online environment, in: Proceedings of the International Conference on Business Excellence, volume 13, Sciendo, 2019, pp. 1108–1117.
[35] B. Ross, J. Heisel, A.-K. Jung, S. Stieglitz, Fake news on social media: The (in)effectiveness of warning messages, 2018. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062513767&partnerID=40&md5=ec93644f2fd7acbd309575f207a11b46, cited by: 32.
[36] Y. Hua, A. Namavari, K. Cheng, M. Naaman, T. Ristenpart, Increasing adversarial uncertainty to scale private similarity testing, in: 31st USENIX Security Symposium (USENIX Security 22), USENIX Association, Boston, MA, 2022, pp. 1777–1794. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/hua.