AI & SOCIETY
https://doi.org/10.1007/s00146-025-02680-5

REVIEW

A toolbox to deal with misinformation in disaster risk management

Rosa Vicari[1] · Nadejda Komendatova[1] · Or Elroy[1,2] · Irina Dallo[3] · Carmit Rapaport[4] · Camille de Carolis[5] · Abraham Yosipof[1,4]

Received: 16 March 2025 / Accepted: 2 October 2025

Abstract
Misinformation significantly challenges disaster risk management by increasing risks and complicating response efforts. This technical note introduces a methodology toolbox designed to help policy makers, decision makers, practitioners, and scientists systematically assess, prevent, and mitigate the risks and impacts of misinformation in disaster scenarios. The methodology consists of eight steps, each offering specific tools and strategies to help address misinformation effectively. The process begins with defining the communication context using PESTEL analysis and Berlo's communication model to assess external factors and information flow. It then focuses on identifying misinformation patterns through data collection and analysis using advanced AI methods. The impact of misinformation on risk perceptions is assessed through established theoretical frameworks, guiding the development of targeted strategies. The methodology includes practical measures for mitigating misinformation, such as implementing AI tools for prebunking and debunking false information. Evaluating the effectiveness of these measures is crucial, and continuous monitoring is recommended to adapt strategies in real-time. Ethical considerations are outlined to ensure compliance with international laws and data privacy regulations. The final step emphasizes managerial aspects, including clear communication and public education, to build trust and promote reliable information sources. This structured approach provides practical insights for enhancing disaster response and reducing the risks associated with misinformation.

Keywords Misinformation · Risk perception · Disaster risk communication · Natural language processing · Artificial intelligence · Ethics

1 Introduction and background

Misinformation during disasters can intensify risks and hinder effective disaster risk management (DRM). This paper introduces a systematic methodology to assess social media misinformation risks and impacts in DRM. By offering structured tools and strategies, it aids researchers, policymakers, decision makers, and practitioners in understanding, preventing, and mitigating misinformation, ultimately fostering more resilient communities and enhancing response efforts.

✉ Rosa Vicari
vicari@iiasa.ac.at

Nadejda Komendatova
komendan@iiasa.ac.at

Or Elroy
ore@uoregon.edu

Irina Dallo
irina.dallo@usys.ethz.ch

Carmit Rapaport
crapaport@geo.haifa.ac.il

Camille de Carolis
camille.decarolis@emsc-csem.org

Abraham Yosipof
avi.yosipof@gmail.com

[1] International Institute for Applied Systems Analysis, Laxenburg, Austria

[2] University of Oregon, Eugene, United States

[3] ETH Zurich, Zurich, Switzerland

[4] College of Law and Business, Ramat Gan, Israel

[5] European-Mediterranean Seismological Centre, Paris, France

Published online: 22 October 2025

1.1 Methodologies and strategies to address misinformation across contexts.

The scientific literature identifies different types of information disorders, among which misinformation is commonly understood as "false" or "misleading" information, shared without the intent to deceive. Lazer et al. (1979) define it in contrast to disinformation, which is deliberately false and spread with the intent to mislead. They place both within the broader context of "fake news," a term they describe as fabricated content mimicking news but lacking journalistic intent or process. Ireton and Posetti (2018a) similarly highlight misinformation and disinformation as core categories of information disorder and caution against the use of "fake news" due to its politicization and its use to discredit journalism. While DiFonzo and Bordia (2007) focus on rumors—unverified and socially meaningful information circulating in uncertain contexts—the present study emphasizes misinformation and rumors as broad, commonly used terms in scholarly work to encompass various forms of misleading or false content, including disinformation and hoaxes.

Misinformation pervades various contexts, prompting diverse methodologies to combat its spread. Lewandowsky et al. (2017) introduced "technocognition", a concept combining cognitive science with technology to design systems that nudge (Thaler and Sunstein 2008) individuals away from misinformation. This approach, complemented by public education and improved journalism, addresses the emotional and belief-driven nature of misinformation in the post-truth era. Similarly, Lazer et al. (1979) called for a multidisciplinary response akin to post-World War II propaganda strategies, incorporating psychology, computer science, political science, economics, law, and communication. The authors highlighted that platform-based interventions, like prioritizing source quality, reducing content personalization, and combating bots, are more effective than individual users' efforts in combating fake news. Conversely, Pennycook and Rand (2019) proposed using crowdsourced trust ratings to refine social media algorithms, as well as increasing visibility for trusted media on social media platforms.

Another interdisciplinary approach was proposed by Wardle and Derakhshan (2017) who explored the social, political, and technical dimensions of misinformation, and emphasized its amplification by platforms that prioritize engagement metrics like likes and shares. Despite fact-checking initiatives, misinformation driven by emotions, like fear and anger, spreads rapidly. Their report proposed 34 recommendations, including advisory councils for tech companies, algorithm transparency, data sharing, filter bubble mitigation, and public education in media literacy. Governments were urged to regulate ad networks, fund public service media, and standardize news literacy curricula. Media organizations were advised to collaborate, maintain ethical standards, and prioritize debunking misinformation.

The literature reviewed above emphasize the importance of integrating multidisciplinary research, psychological principles, AI tools, platform-based solutions, and community-driven trust ratings. In the next section, we will review literature focusing on the context of disasters, where misinformation poses unique challenges and requires tailored strategies.

1.2 Misinformation in disasters: insights and strategies

The proliferation of misinformation during disasters poses significant challenges for effective risk management and communication. According to Wisner al. (2004) "A disaster is the result of a hazard's impact on society. So the effects of a disaster are determined by the extent of a community's vulnerability to the hazard and the effectiveness of measures to reduce or cope with the potential harmful effects". Various studies have explored the dynamics of fake news spread during disasters and the strategies to counteract it. Oh et al. (2013) applied rumor theory to study collective reporting on Twitter during social crises, such as the Mumbai terrorist attacks in 2008 and the Toyota recall in 2010. They found that unclear sources, personal involvement, and anxiety drove rumor propagation, underscoring the need for transparent and authoritative information sources to mitigate misinformation risks during crises.

Building on the importance of fact-based information, Paek and Hove (2019) analyzed strategies for countering rumors about radiation-contaminated food in South Korea and identified three main tactics: refuting the rumor with facts and evidence, outright denial without evidence, and attacking the source of the rumor. Their findings highlight the superior effectiveness of evidence-based refutation in reducing misinformation spread. Similarly, Hunt et al. (2020) analyzed false rumors during Hurricanes Harvey and Irma, demonstrating that authoritative sources, such as verified government accounts, debunk misinformation effectively. URLs and news agencies played key roles in countering false narratives.

During the COVID-19 pandemic, misinformation dynamics prompted novel strategies. Papakyriakopoulos et al. (2020) found that mainstream sources contribute more significantly to the spread of conspiracy theories on social media compared to alternative sources. They observed that while content moderation helps curb misinformation, challenges persist in ensuring timeliness, effectiveness, and transparency, highlighting the need for clear communication and deliberation of content removal to build user trust and awareness on social media. This insight aligns with Paek and Hove's (2019) emphasis on evidence-based refutation, suggesting that transparency and clear communication are key components of effective misinformation management. Pian et al. (2021) also highlighted the importance of audience-tailored risk communication to combat the COVID-19 infodemic.

Various studies focus on AI tools to detect and combat the COVID-19 infodemic. Salehinejad et al. (2021) advocated for the development of automated, real-time tools to detect rumors during crises, which are critical in fast-paced social media environments. Varma et al. (2021) reviewed fake news detection technologies, highlighting the potential of deep learning algorithms for high-accuracy misinformation identification. This technical perspective complements Micaleff et al. (2020) who observed how swiftly users responded to pandemic-related misinformation on Twitter. The authors recommended developing tools to empower citizens to combat misinformation, highlighting the role of user engagement and technology in addressing fake news.

Naeem and Boulos (2021) advocate for synergistic strategies that combine machine learning with fact-checking, involving both content and source evaluation. This comprehensive approach resonates with Liu and Xiao's (2021) call for integrating health literacy education, digital literacy education, and Internet access to improve eHealth literacy. Both groups of authors emphasized the need for education and technical innovation to work hand in hand.

In summary, recent literature highlights that addressing fake news during disasters requires a combination of fact-based refutation, transparent content moderation, authoritative sources, education, and advanced technologies. While numerous methodologies and strategies exist, no general and comprehensive framework currently addresses and manages misinformation related to both anthropogenic and natural hazards and disasters. This research paper aims to develop and validate a methodological framework for addressing various forms of misinformation relevant to disaster risk reduction.

1.3 Objectives

This toolbox aims to provide a methodological framework for addressing various forms of misinformation relevant to disaster risk reduction. The methodology consists of eight steps, as shown in Fig. 1, addressing communication patterns, influence of misinformation on risk perceptions, ethical challenges, stakeholder preferences for misinformation-fighting tools. For each methodological step, we provide case studies—either from the authors' own research or from the existing literature—that demonstrate its application. Cognitive and behavioral biases influencing risk perception and awareness are considered, alongside the interdependence between risk perception, awareness, and motivation for action.

The development of this framework was informed by an extensive review of the existing academic literature on

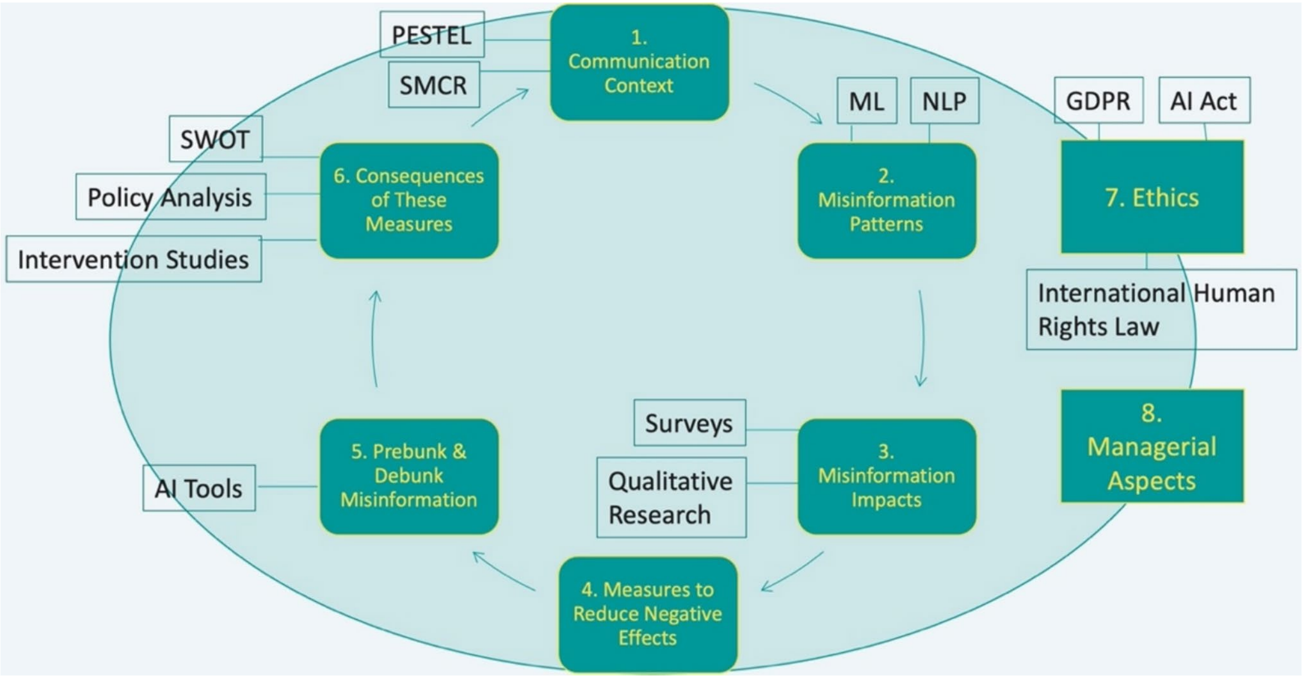

PESTEL
SMCR
1. Communication Context
ML
NLP
GDPR
AI Act
SWOT
Policy Analysis
Intervention Studies
6. Consequences of These Measures
2. Misinformation Patterns
7. Ethics
International Human Rights Law
Surveys
Qualitative Research
3. Misinformation Impacts
8. Managerial Aspects
AI Tools
5. Prebunk & Debunk Misinformation
4. Measures to Reduce Negative Effects

Fig. 1 Eight steps to tackle misinformation in disaster risk management

misinformation, disaster communication, risk perception, and behavioral science. Tools were selected for inclusion based on their empirical validation, relevance to disaster contexts, applicability across different types of hazards, and demonstrated effectiveness in previous case studies. We prioritized tools that offered flexibility and adaptability to various contexts, ensuring the framework's broad relevance.

The framework involved a co-creation approach, involving two groups of social media users. The first one is a broader public whose reaction on AI tools helped us to formulate the recommendations needed for this framework. The second group are disaster risk reduction stakeholders, first responders and the users of the EMSC (European-Mediterranean Seismological Centre) prevention and debunking tools in social media. We engaged disaster risk reduction stakeholders through surveys and focus group discussions to assess their perceptions of AI tools for combating misinformation about hazards and to refine the methodological framework based on their feedback, as detailed in Sect. 5.1.

These steps consist in a selection of tools aimed at providing a comprehensive approach to address misinformation. However, it is important to note that these tools and steps are not necessarily intended to be implemented exhaustively. Instead, they should be chosen and adapted according to the specific context of implementation, ensuring that the most relevant and effective strategies are applied in each unique situation.

The literature defines various information disorders. Lazer et al. (1979) describe fake news as fabricated content mimicking news but lacking journalistic intent. Ireton and Posetti (2018b) prefer 'misinformation' and 'disinformation,' rejecting 'fake news' due to its politicized use to discredit journalism. For this study, we adopt the broader terms 'misinformation' and 'rumors' to encompass a range of disorders, including disinformation and hoaxes. Additionally, we use 'conspiracy theory' to refer to false narratives attributing events to a malevolent scheme by multiple actors.

2 STEP 1: Define the communication context

As the initial step in developing an effective methodology to address misinformation in the context of disaster risk reduction, it is crucial to define the communication context. Understanding the communication context is essential because it sets the stage for identifying the unique challenges and opportunities presented by different types of hazards and disasters, whether anthropogenic or natural. It allows for a tailored approach that considers the specific characteristics of the information environment, the behavior of audiences, and the nature of the misinformation being spread. By clearly defining the communication context, researchers and practitioners can develop more targeted strategies to combat misinformation, ensuring that interventions are relevant and effective for the disaster scenario at hand. This step helps to build a robust framework that enhances public understanding, mitigates misinformation risks, and supports informed decision-making during critical times.

2.1 PESTEL analysis to identify the political, economic, social, technological, environmental, and legal factors

PESTEL analysis (Aguilar 1967; Nandonde 2019) is a strategic tool used to assess the external macro-environmental factors affecting organizations or projects. The acronym PESTEL stands for Political, Economic, Social, Technological, Environmental, and Legal factors. PESTEL provides a holistic view of the external factors that could influence communication strategies during a disaster. For example, Kung (2023) employed the PESTEL framework to systematically assess the effectiveness of emerging digital media strategies, such as live streaming, NFTs, and the Metaverse. By analyzing the political, economic, social, technological, environmental, and legal dimensions, the study offers insights into how these macro-environmental factors shape the adoption, implementation, and overall success of communication platforms and strategies. Table 1 presents examples of variables for each of the six PESTEL categories. Understanding these

Table 1 PESTEL analysis with examples (Dubetcky 2024)

P	E	S	T	E	L
• Government policy • Political stability • Corruption • Foreign trade policy • Labor law	• Economic growth • Inflation rates • Disposable income • Unemploy-ment rates	• Population growth rate • Age distribution • Safety emphasis • Health consciousness • Lifestyle attitudes • Cultural barriers	• Technology incentives • Level of innovation • Automation • R&D activity • Technological change • Technological awareness	• Weather • Climate • Environmental policies • Climate change • Pressures from NGOs	• Discrimination laws • Employment laws • Consumer protection laws • Copyright and patent laws • Health and safety laws

factors helps in formulating effective and context-specific strategies to tackle misinformation.

Political factors involve government policies, stability, and interventions that may impact operations or strategies. In a disaster scenario, government policies and regulations can greatly influence communication. Knowing the political landscape helps in aligning messages with official guidelines.

Economic factors encompass growth, exchange rates, and economic conditions shaping decision-making processes. Economic conditions can affect resource availability and the public's response to communication efforts. Economic stability or instability will shape how resources are allocated for communication and how messages are received by different economic groups.

Social factors consider societal norms, cultural backgrounds, demographics, and lifestyle changes. Social dynamics, such as public sentiment, cultural norms, and community structures, are crucial in tailoring messages that resonate with different demographics. Understanding these factors ensures the communication is culturally sensitive and socially appropriate.

Technological factors cover technological advancements, innovation, research, and development. Analyzing technological factors helps in selecting the most efficient channels for accurate information dissemination and anticipating potential technological opportunities and challenges. This includes understanding the capabilities of various social media platforms, access challenges to these platforms during disasters and emergencies, the role of algorithms in content visibility, and the effectiveness of existing tools for fact-checking and monitoring the spread of false information.

Environmental factors encompass environmental and ecological aspects such as climate change, tensions on natural resources availability, ecosystem protection, environmental regulations, and sustainability issues. Environmental factors play a significant role in shaping the communication context during natural or anthropogenic disasters or hazards. The type and severity of the hazard or disaster itself are environmental factors that shape the communication context. Different hazards, such as wildfires, chemical spills, or infectious disease outbreaks, require tailored communication strategies to address specific risks and inform appropriate responses.

The physical environment, including geology, weather conditions, and infrastructure, can impact communication capabilities. The geographical location of the disaster or hazard can influence the communication context. Coastal areas prone to hurricanes or low-lying regions susceptible to flooding may have different communication needs compared to urban areas facing industrial accidents or technological hazards.

Legal factors pertain to laws and regulations affecting operational or strategic aspects. Legal considerations are important for ensuring that the communication complies with laws and regulations. This includes understanding privacy laws, freedom of information acts, and other legal constraints that might affect how information is shared.

2.2 Source, message, channel, receiver, feedback

Analyzing the general communication context of anthropogenic or natural hazards and disasters is crucial before addressing any misinformation issue. Understanding this context provides essential insights into how information flows, is received, and impacts decision-making processes during emergencies. Both qualitative and quantitative methods, including informal feedback, are essential for apprehending the communication context. Berlo's communication model, introduced in 1960, outlines the key steps of communication: Source, Message, Channel, Receiver, Effect, and Feedback (Dallo et al. 2023). Berlo's model was selected because it offers a clear, systematic framework for analyzing the key components of communication—source, message, channel, and receiver—which are central to our methodology focus on risk and crisis communication. Its emphasis on how audience characteristics, message content, and channel choice interact aligns directly with our aim to tailor communication strategies to diverse audiences, especially in contexts of misinformation and disaster management. Compared to other models, Berlo's approach offers practical analytical clarity for understanding how communication effectiveness can be optimized in situations of uncertainty, as demonstrated in Dallo's thesis (2022), which applies Berlo's communication model to examine the full communication chain for earthquake-related information—from message source, design, and channels to public reception, understanding, action, and iterative evaluation—thereby highlighting the complexity and interdependence of each stage in a multi-hazard context.

2.2.1 Sources

Identifying the source of information is essential to understand where information originates during disaster events. This includes identifying official sources such as government agencies, emergency services, and credible media outlets, as well as unofficial sources like social media users or citizen journalists. Nowadays, trust in information sources has become crucial, with authorities and experts often being most trusted during emergencies.

2.2.2 Messages

Messages, particularly during crises, should contain essential elements such as the hazard type, affected area, time, source, and behavioral recommendations. Understanding the message involves analyzing the accuracy, reliability, and potential biases of the information, its format, as well as its relevance to the disaster situation.

2.2.3 Channels

Mediums used for communication include traditional media such as television, radio, and newspapers, as well as digital platforms like social media, websites, and mobile apps. Different channels have varied reach and accessibility to different segments of the population, with preferences often influenced by factors such as age, geographic location, or cultural background. Some channels facilitate rapid dissemination of information, while others may experience inherent delays. The choice of communication channels can impact the perceived trustworthiness and credibility of the information being shared. Social media, for instance, play a significant role in disseminating information rapidly and facilitating two-way communication, although they also pose risks of misinformation spread.

2.2.4 Receivers

Personal and contextual factors influence how receivers interpret and respond to messages, as demonstrated by various social cognition models. Analyzing the receiver includes considering factors such as demographics, literacy levels, cultural backgrounds, social, and physical environment.

2.2.5 Effects and feedback

Various factors influence how receivers respond to a message (e.g., self-efficacy, knowledge, and prior experiences). Predicting those factors can be used to design effective information campaigns or warning messages. Evaluating feedback allows for adjustments in the communication process to enhance effectiveness, considering changing needs and technological advancements. Both qualitative (focus groups, interviews, content analysis) and quantitative (surveys, web analytics, and behavioral data) methods are essential to apprehend the feedback loop and the effectiveness of the entire communication process.

Tracking searches on Wikipedia can reveal public interests, risk perceptions, and changing preferences over time. Yosipof and Rapaport (2023), compared Wikipedia page traffic data in multiple languages for six key case studies: the L'Aquila earthquake, Manchester Arena bombing, Aude River flooding, Visakhapatnam gas leak, Tōhoku earthquake and tsunami, the COVID-19 pandemic, and the 2021 European floods. For instance, analysis of the Wikipedia page for the Manchester Arena attack reveals significant patterns (Yosipof et al. 2023): regular peaks on Memorial Day (May 22), with the highest peak on the first anniversary due to heightened media coverage. Additional notable peaks occurred on August 22, 2020, for Hashem Abedi's conviction, and on November 3, 2022, following the release of a key public inquiry report on the emergency response.

Defining the communication context is a foundational step that ensures the strategies developed are well aligned with the unique challenges posed by different disaster scenarios. With a clear understanding of the external factors and communication dynamics, we can move forward to examining the misinformation patterns related to a disaster with specific data sources and analytical techniques.

3 STEP 2: Identify current misinformation patterns

Building on the understanding of the communication context, the next step involves identifying the patterns of misinformation that typically arise during and after disaster scenarios. Recognizing these patterns is essential for developing effective strategies to counter misinformation and ensure accurate information dissemination. In this session, we will examine the sources of data and methods used to detect and analyze misinformation patterns.

3.1 Sources of data

In the context of disaster response, social media, press coverage, surveys, and official documents from public authorities serve as sources of data for identifying patterns of misinformation. In times of disaster, these sources are particularly relevant as they provide real-time information and public discourse snapshots, enabling a comprehensive understanding of misinformation dynamics and facilitating prompt intervention to mitigate its impact on public safety and decision-making processes.

3.1.1 Social media

Social media platforms such as X (formerly Twitter), Facebook, and Instagram can be instrumental in monitoring misinformation during hazards or disasters. X's real-time updates and widespread use make it an effective tool for tracking emerging rumors and false information, as users often share immediate reactions and unverified reports. Facebook, with its extensive user base and community groups, can be used to identify and counteract misinformation spreading within local communities and networks. Instagram, while more visual, can help detect misleading images and videos that might be circulating. Additionally, platforms like WhatsApp and Telegram, which are popular for private messaging, can also be monitored for misinformation that spreads through personal networks. Utilizing these platforms allows authorities and fact checkers to quickly identify, address, and correct false information, thereby helping to maintain public safety and trust during critical events**.**

Public authorities and researchers can access social media data through a combination of official APIs, third-party tools, and partnerships with the platforms themselves. Many social media platforms, such as X and Facebook, offer APIs that provide access to a wide range of data, including posts, user interactions, and trending topics. Researchers can use these APIs to collect and analyze data in real-time, allowing them to monitor misinformation and public sentiment during hazards or disasters. Additionally, third-party analytics tools and services can offer sophisticated data collection and analysis capabilities, often aggregating data from multiple platforms. Partnerships with social media companies can also facilitate access to data, especially during emergencies when rapid and comprehensive data collection is crucial.

For instance, tweets extracted from the X API were used in four studies, each focusing on a specific case of misinformation on X: (i) the COVID-19 conspiracy theories tweets (Erokhin et al. 2022; Elroy and Yosipof 2022), (ii) the Monkeypox tweets (Elroy et al. 2023), (iii) the earthquake prediction tweets (Elroy and Yosipof 2023; Dallo et al. 2023), and (iv) five rumors regarding the Manchester Arena attack (Vicari et al. 2024).

The authors used the X API's v2 full search endpoint, tailored for Academic Research, which provides access to the entire X archive via a key terms search query. This query retrieves tweets that match the specified criteria from the complete archive, together with metadata of the tweets and the authors (such as the date and time of publication, the number of followers, the number of followings, if the tweet contains an URL and the number of retweets). Retweets were omitted from the searches, and only tweets identified as English by Twitter's language detection algorithm were included. Each database was curated for a specific time frame.

Press: Analysis of press news related to a hazard or disaster provides valuable insights into the dissemination and evolution of misinformation. By examining news articles, it is possible to identify prevalent rumors, understand the prevalent themes and narratives, assess their impact on public perception, and track the spread of false narratives over time. Additionally, analyzing press coverage allows for the identification of key misinformation sources, such as social media platforms.

Public authorities and researchers can access press news data through various means, including online news archives, media-monitoring services, and partnerships with news agencies. Many news organizations offer digital archives or APIs that grant access to their articles and metadata. Additionally, media monitoring services aggregate press coverage from multiple sources, providing comprehensive datasets for analysis. Collaboration with news agencies can also facilitate access to proprietary data and insights.

Vicari et al. (2024) curated a dataset consisting of English press articles concerning the Manchester Arena attack. These articles were sourced from Europresse's database, encompassing 439 global media outlets (Cision 2023). The search parameters focused on specific keywords present solely in article titles and spanned 6 years, starting from the day of the attack. Gugg (2024) discussed the Mount Vesuvius case study, which featured an analysis of 130 Italian press articles about Vesuvius published between 2012 and 2022 that extensively circulated both online and offline.

Surveys: Surveys allow researchers to directly query individuals about their beliefs, perceptions, and sources of information during crisis events. Surveys can uncover the prevalence of misinformation and identify common misconceptions. To carry out a survey, researchers typically design questionnaires tailored to the specific context of the disaster, incorporating questions about individuals' exposure to and beliefs regarding misinformation. Surveys can be administered through various methods such as online platforms, telephone interviews, or in-person interactions, ensuring a diverse and representative sample population for comprehensive analysis (Jensen and Laurie 2016). Surveys can complement other sources of data, such as social media and the press. Integrating survey data with information from these sources allows for triangulation and validation of findings, enhancing the robustness of insights gained.

Interviews and focus groups: These qualitative methods allow for in-depth exploration of individuals' experiences, beliefs, and perceptions regarding misinformation in the context of crisis events. Through interviews, researchers can delve into participants' thought processes, motivations, and information-seeking behaviors, uncovering nuanced insights that may not be captured through quantitative surveys alone. Similarly, focus groups facilitate dynamic discussions among participants, allowing for the exploration of diverse perspectives and the identification of common themes and misconceptions. To carry out interviews and focus groups, researchers typically develop semi-structured interview guides or discussion protocols tailored to the research objectives and participant demographics (Jensen and Laurie 2016). These sessions can be conducted in person or remotely, ensuring flexibility and accessibility for participants. The data collected from interviews and focus groups complement quantitative findings, providing a richer understanding of the complexities surrounding misinformation during disasters.

3.2 Methods of analysis

In examining methods to detect misinformation patterns in disaster contexts, a plethora of qualitative and quantitative approaches exist. However, our focus lies on recent techniques developed for analyzing misinformation on social media, supported by automated tools, with the aim of facilitating rapid responses. These methods offer valuable insights into identifying and addressing misinformation during crises.

As displayed in Fig. 2, the dataset phase consists of collecting the appropriate data, manually classifying a subset of the dataset into different groups and embedding the semantic meaning of the textual features in the dataset using Natural Language Processing (NLP) methods. Next, a classification methodology needs to be designed to classify the complete dataset into the relevant groups based on the subset that was manually classified. Finally, a thorough analysis must be performed on the labeled dataset to produce valuable insights and recommendations. To enhance transparency and reproducibility, supplementary material has been added. It provides detailed descriptions of the datasets associated with the case studies referenced in in this section, including their sources, structure, and scope of application.

3.2.1 Natural language processing methods

A primary challenge in analyzing misinformation in texts is reliably classifying them. Advancements in Natural Language Processing (NLP) have introduced new algorithms for text embedding and classification, such as Bidirectional Encoder Representations from Transformers (BERT). Elroy and Yosipof (2022) used the Covid-Twitter-BERT model that was pre-trained on COVID-19-related tweets to classify tweets as supporting or opposing the COVID-19 5G conspiracy.

Elroy et al. (2023) and Dallo et al. (2023) used the Robustly Optimized BERT Pretraining Approach (RoBERTa), which is an enhanced BERT with different design decisions resulting in improved performance. In the first study, the authors categorized the entire dataset on the Monkeypox outbreak into three groups: misinformation, counter-misinformation, and neutral. The second study is based on a classification of tweets as either misinformation or not misinformation regarding earthquakes.

3.2.2 Machine learning

Supervised learning models train on manually labeled samples from a dataset by associating their features as input and predicting their label as output. Therefore, developing a successful supervised learning classifier depends on large amount of labeled data for training, which is traditionally obtained through an intensive process of manual labeling. A too small subset of labeled samples may present different issues in the classification model.

Semi-supervised learning (SSL) tries to resolve the necessity of large amounts of manually labeled data by enriching the labeled dataset with pseudo-labeled samples derived from assumptions about suitable labels for some unlabeled data. At its basic form, a supervised learning model is trained on the small subset of labeled data and used to predict the labels for all other samples, where the predictions with a high certainty are assumed to be pseudo-labels. More sophisticated variations exist, such as adjusted semi-supervised learning for social media (ASSLSM) introduced by Elroy and Yosipof (2023). Evaluations and comparisons of different approaches showed that the ASSLSM method improved SSL's pseudo-labeling process, resulting in more consistent performance across various classifiers other than standard SSL.

3.2.3 Natural language processing and machine learning model performance

Elroy and Yosipof (2022) used an ensemble of classifiers combining sentence embeddings from COVID-Twitter-BERT and Sentence-BERT with external features like number of followers and average sentiment scores, to classify tweets related to the COVID-19 5G conspiracy theory, achieving an F1 score of 0.904. For Monkeypox misinformation, a RoBERTa model fine-tuned on 3218 hand-labeled

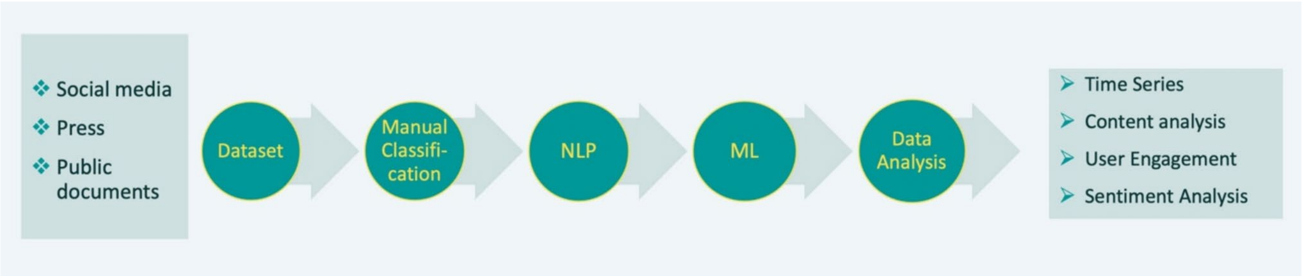


Fig. 2 Workflow to analyze misinformation of social media with the support of automated tools

tweets and was used to classify 1.4 M tweets, achieving an average F1 score of 0.767, with precision and recall both at 0.774, using stratified fivefold cross-validation (Elroy et al. 2023). In the context of earthquake prediction misinformation Dallo et al. (2023) construct a classifier using RoBERTa embedding to achieve an F1 score of 0.845. The semi-supervised learning approach, namely adjusted semi-supervised learning for social media (ASSLSM), improved the performance of earthquake misinformation further, reaching F1 scores up to 0.969 with ensemble classifiers (Elroy and Yosipof 2023). Vicari et al. (2024; Elroy and Yosipof 2022, 2023; Elroy et al. 2023; Dallo et al. 2023) examined five rumors related to the Manchester Arena bombing using tweet classification and engagement analysis over a six-year period to assess the persistence and emotional impact of misinformation. Although these studies demonstrate the effectiveness of AI-based classification models, they also underscore limitations related to the dependence on manually labeled datasets, language constraints, and potential domain-specific generalization issues.

3.2.4 Time series

A descriptive analysis of the frequency of misinformation and accurate information, or various categories of misinformation over time, can reveal temporal patterns and influencing factors. For instance, Dallo et al. (2023) discovered that daily peaks in misinformation and accurate information often correlate, indicating that the spread of misinformation related to earthquake predictions increases after significant events and during earthquake sequences. An Ordinary Least Squares (OLS) time series model was used to analyze the effect of accurate information tweets on the spread of misinformation tweets and vice versa. The results revealed that the continuous presence of earthquake predictions on X indicates a constant need for communication strategies to counteract them, even during periods of low seismic activity.

Another example is provided by Erokhin et al. (2022) who represented each COVID-19 conspiracy theory as a time series of tweet frequency. They performed correlation and cross-correlation analysis between the conspiracy tweet frequencies to identify dependencies among different conspiracy discussions. As a result, they classified conspiracy theories into four categories: (1) those that reached their highest levels early in the pandemic (like 5G and FilmYourHospital), (2) those that gained traction as time passes (such as Big Pharma and vaccination theories), (3) those that persisted throughout (such as exaggeration and Bill Gates theories), and (4) those that experienced multiple peaks (like GMO and biological weapons theories). Additionally, they applied an OLS time series model and found that, for most of these conspiracy theories, the frequency of tweets in the first week after a theory's emergence was significantly linked to the number of new COVID-19 cases.

Elroy et al. (2023) examined the weekly frequency time series of tweets spreading Monkeypox misinformation and those countering it. They observed that tweets spreading misinformation about Monkeypox consistently dominated the discussion since the outbreak's onset. However, a shift in dominance from misinformation to accurate information occurred at the start of the second peak of tweets related to Monkeypox, indicating the final phase of the misinformation spread.

3.2.5 Content analysis

Content analysis is a research method used to systematically analyze communication contents from various media sources. By coding and categorizing texts, images, URL or audio-visual materials, researchers can identify frequency and context of specific themes, terms, narratives, and misinformation types.

Thanks to content analysis, Dallo et al. (2023) observed that X users often link earthquake notifications from official sources in their earthquake prediction claims. This is particularly concerning because readers may trust the official source and, as a result, also believe the misinformation. Consequently, public authorities and agencies should regularly ensure that their notifications are not linked to, nor their accounts tagged in, misinformation tweets. Elroy et al. (2023) highlighted that tweets containing misinformation related to the Monkeypox outbreak mostly reference domains where users can upload and publish content themselves, such as YouTube, followed by other websites that often refer to extreme free speech and conspiracy theories.

3.2.6 User engagement

The interactive nature of social media offers a significant advantage in understanding how misinformation about disasters is perceived by the public. Engagement metrics such as likes, shares, comments, and replies are crucial indicators of a post's resonance on social media platforms. A high level of likes and shares reflects the content's importance, appeal, and relevance to a wide audience. Posts with substantial engagement often reach a larger audience, increasing visibility through sharing. The number of shares serves as an essential measure of virality, demonstrating that the content has captured the attention of a broad and growing audience.

Engagement metrics were explored by Dallo et al. (2023), who compared reactions to misinformation tweets, versus not-misinformation tweets related to earthquakes. They found that the mean number of retweets, likes, replies in response to not-misinformation tweets was lower than the mean number of retweets, likes, replies in response to misinformation tweets.

Vicari et al. (2024) used tweet engagement to compare five rumors related to the Manchester Arena attack, analyzing hourly occurrences over 2 days post-attack and monthly frequencies over 6 years. A rumor about the attacker being a refugee elicited the most significant response. Short-term impacts were noted for rumors about children in hotels and a gunman in Oldham Hospital, while the refugee rumor gained traction over 6 years, likely influenced by anti-immigrant sentiments exacerbated by political narratives, following the 2019 Boris Johnson election campaign.

3.2.7 Sentiment analysis

Sentiment analysis is a technique used to determine the emotional tone behind a body of text. It provides insights into public opinion and reactions, helping to understand how misinformation spreads, it influences perceptions, and affects emotions during disasters or crises.

The VADER sentiment analysis model (Hutto and Gilbert 2014) is particularly useful for understanding textual nuances in informal communication, like tweets. Unlike traditional sentiment analysis models that focus on individual words and their polarities, VADER considers the context of words within a sentence, including the effects of punctuation and capitalization. This enhanced approach captures the subtleties of casual online language. Sentiment analysis helps identify the overall sentiment of the content and measure the emotional intensity associated with it.

In their study on misinformation related to the Manchester Arena attack, Vicari et al. (2024) noted that rumors elicited varied emotional responses from X users, ranging from neutrality to heightened levels of distress, anger, or discontent, depending on the type of rumor. The findings reveal a dual impact of rumors post-attack: initially disrupting factual information and emergency management within the first 2 days, then shaping opinions and emotions in the long term, potentially exacerbating divisions in public sentiment.

Recognizing misinformation patterns related to disasters is essential for understanding how false information spreads and can be addressed. By leveraging data from social media, press coverage, surveys, and other sources, and employing advanced analytical methods, we can track the spread of misinformation and its evolution. In the following session, we will explore how these patterns shape public perceptions of risk and examine the consequential effects on risk management practices during crises.

4 STEP 3: Assess misinformation impact on risk perceptions and risk management

Misinformation significantly impacts the perception of risk among various stakeholders, often amplifying fear and anxiety during crises and health emergencies. This heightened state of concern can lead to widespread panic and hinder effective response efforts. Additionally, misinformation affects decision-making processes in risk management by introducing cognitive biases that distort perception (Dallo et al. 2023). *Anchoring bias* causes reliance on specific information traits, while *availability bias* leads to underestimating low-probability risks until experienced. *Confirmation bias* favors information aligning with existing beliefs, and the *cry-wolf syndrome* erodes trust after repeated false alarms. Other biases, such as *loss aversion* and *myopia*, contribute to inconsistencies in decision-making. Understanding these impacts is crucial for developing strategies that mitigate the influence of misinformation and improve risk management practices.

To examine how misinformation affects risk perceptions and decision-making among various stakeholders, the first step is to develop a conceptual and theoretical framework. This can be done using models like the Theory of Planned Behavior (TPB) (Pundir et al. 2021), the Risk Information Seeking and Processing (RISP) model (Griffin et al. 2004), and the heuristic framework developed by Hasson et al. (2020), as presented in Table 2.

Table 2 Models used to examine how misinformation influences perceptions of risk and decision-making among different stakeholders

Model	Independent variables	Dependent variables
Theory of Planned Behavior (TPB) (Pundir et al. 2021)	• Awareness of fake news Attitudes toward news verification • Perceived behavioral control • Subjective norms • Fear of missing out • Sadism	Social media users' intentions to verify news before sharing it
Risk Information Seeking and Processing (RISP) (Griffin et al. 2004)	• Attributes of risks • Individual characteristics	'Information sufficiency'
Heuristic framework by Hansson et al. (2020)	• Communication-related factors • Individual, social-structural, and situational vulnerabilities	Individuals' abilities to prepare for and respond to disasters

Pundir et al. (2021) applied TPB to explore social media users' intentions to verify news before sharing it. Their study considered factors such as awareness of fake news, attitudes toward news verification, perceived behavioral control, subjective norms, fear of missing out (FoMO), and sadism, highlighting the multifaceted nature of information verification behaviors.

The Risk Information Seeking and Processing (RISP) model, employed by Griffin et al. (2004), investigates how the attributes of risks and individual characteristics influence 'information sufficiency'—an individual's assessment of the necessary information to manage risks effectively. This model underscores the importance of understanding personal and contextual factors in risk information processing.

Another example of a theoretical framework is the heuristic framework introduced by Hansson et al. (2020) that elucidates how communication-related factors can hinder individuals' abilities to prepare for and respond to disasters. This framework aids researchers, policymakers, and practitioners in systematically identifying individual, social-structural, and situational vulnerabilities, thereby enhancing disaster and crisis management strategies.

After establishing the theoretical framework, the following step involves collecting and analyzing data using various qualitative and quantitative techniques, which can be applied alternatively or can be combined.

4.1 Surveys and questionnaires

Surveys and questionnaires are tools for gathering quantitative data on individuals' exposure to misinformation and their perceived risks. For example, Griffin et al. (2004) conducted a telephone survey with 1123 adult residents in Great Lakes cities to understand how individuals use information in risky situations. The advantage of this technique lies in its ability to collect large amounts of data efficiently, allowing for statistical analysis and the identification of patterns and correlations within the population. The survey methodology can follow a structured design that includes carefully formulated questions aligned with the theoretical framework, using both closed- and open-ended items. Respondents can be selected through purposive or stratified sampling to ensure representation across relevant stakeholder groups. Data collection methods (e.g., online surveys, phone interviews, or paper-based questionnaires) should be chosen based on the target population and context, with attention to minimizing bias and ensuring reliability. Pre-testing or piloting the survey instruments helps refine the questions for clarity and relevance before full deployment.

4.2 Experiments

Conducting experiments allows researchers to observe changes in risk perception following exposure to misinformation. Spence et al. (2016) performed a three-condition experiment with 258 participants to study how the speed of updates on Twitter (immediate, recent, and delayed) affects credibility perceptions and information-seeking behavior. Such experimental designs help isolate specific factors influencing risk perceptions and decision-making.

4.3 Field observation

Field observation, a form of field research, involves observing individuals or groups in their natural environment to gain insights into their behaviors, activities, and processes. Herztum et al. (2002) used a field study during the early stages of a major software engineering project to examine how trust influences individuals' information-seeking behaviors concerning people and document sources. This method provides a contextual understanding of how misinformation impacts real-world settings.

4.4 Focus groups

Facilitating focus group discussions enables researchers to gather diverse insights on the impacts of misinformation and allows participants to build on each other's ideas. In their study, Herztum et al. (2002) conducted focus group investigations into individuals' perceptions of virtual agents on e-commerce websites, examining factors that influence trust and information-seeking behaviors. Focus groups offer a rich qualitative perspective on collective attitudes and beliefs.

4.5 Interviews

In-depth interviews with individuals provide a deeper understanding of their perceptions and decision-making processes. Das et al. (2022) combined Twitter data analysis with in-depth interviews to explore ideological factors influencing citizens' acceptance or rejection of disinformation during the COVID-19 pandemic. Interviews offer nuanced insights into personal experiences and rationales behind behaviors.

In the case study on Mount Vesuvius presented by Dallo et al. (2023), the social effects of the articles and their language were examined using data gathered through ethnographic methods (field interviews) and ethnographic observation (systematic monitoring of online discussions related to the most widely shared articles).

4.6 Delphi method

The Delphi method involves a group of experts providing data anonymously through multiple questionnaire rounds to achieve consensus on complex topics. Flostrand et al. (2019) used the Delphi method to consolidate the perspectives of 42 brand management academics on the perceived threat level, engagement, and actionable strategies for brand managers in response to fake news. This method ensures a well-rounded and expert-informed understanding of issues.

Understanding how misinformation distorts risk perceptions and decision-making is essential for addressing its detrimental effects on public responses during crises. By examining these impacts through diverse research methods, we gain a clearer understanding of the cognitive biases that amplify the influence of misinformation and how to effectively counteract them. In the next session, we will focus on practical measures that institutions, policy makers, decision makers, practitioners, and scientists can implement to mitigate the negative effects of misinformation on risk management efforts.

5 STEP 4: Implement measures to mitigate negative effects

Dallo et al. (2023) provide recommendations for preventing and combating the spread of misinformation across the entire communication chain, encompassing the source, message, channel, receiver, effect, and feedback. These recommendations are derived from six case studies, presented in Table 3, addressing various hazards and time periods.

This approach ensures that the recommendations are aligned with the current dynamics of social media, making them particularly relevant for public authorities and institutions seeking to counteract misinformation effectively today. The recommendations are valid for natural and anthropogenic hazards as well as multi-hazard contexts.

a. *Source identification and trust building*: Source identification and trust building involve identifying and prioritizing trusted sources such as official authorities, press agencies, and scientific experts. It is essential to invest

Table 3 *Source, method, sample size, period, and hazard focus of the six case studies* (Dallo, et al. 2023)

Case studies	Data source	Method	Sample size	Period	Hazard
Misinformation about the link between Covid-19 and 5G on Twitter (Elroy and Yosipof 2022)	Twitter [English]	• Natural Language Processing methods • RoBERTa • Quantitative analysis	$N=331{,}448$	01/01/2020–12/31/2021	Pandemic
Misinformation about earthquake predictions on social media (Elroy and Yosipof 2023; Dallo et al. 2023)	Twitter [English]	• Natural Language Processing methods • RoBERTa • Ordinary least squares time series model • Quantitative and qualitative analysis	$N=82{,}129$	03/01/2020–03/31/2022	Earthquake
Fake news about the volcano Vesuvius on general news media (Gugg 2024)	Local online news media [Italian]	• Media analysis • Interviews	$N=130$ articles	2012–2022	Volcano
Mining the discussion of Monkeypox misinformation on Twitter (Elroy et al. 2023)	Twitter [English]	• Natural Language Processing methods • RoBERTa • Quantitative analysis	$N=1{,}440{,}475$	05/01/2022–08/24/2022	Epidemic
Misinformation and the role of media after the Manchester Arena attack (Vicari et al. 2024)	Twitter, worldwide press [English]	• Natural Language Processing methods • Descriptive statistics • Sentiment analysis	3505 press articles 89,147 tweets	05/22/2017–03/13/2023	Terrorist attack
Authoritative policies to increase societies' resilience to earthquakes—a cross-cultural comparison (Rapaport et al. 2024)	Authoritative documents, nationwide public surveys [Israeli, German]	• Descriptive case study comparison	Surveys CH: $N=596$ IL: $N=920$	2020–2023	Earthquake

in building trust in these sources within society to ensure that their information is perceived as reliable.

b. *Message tone and content*: Official messages should maintain an objective and neutral tone, providing context for the information. To counter the dramatic and attractive elements of misinformation, it is essential to present facts calmly and clearly.
c. *Channel management and cross-verification*: Monitoring and countering misinformation across various social media platforms and communication networks can significantly reduce the spread of false information. Additionally, encouraging the cross-verification of information across different channels helps to mitigate confirmation biases, ensuring that individuals do not rely solely on a single source and are more likely to encounter accurate information.
d. *Tailored strategies for different receiver groups*: To address the diverse groups of receivers, it is essential to develop tailored strategies that cater to their specific needs and behaviors. These strategies should consider those who are actively seeking information, those who passively encounter misinformation, and followers of specific accounts. Different approaches must be employed to persuade individuals involved in conspiracy networks compared to those who are incidentally exposed to misinformation.
e. *Consider emotional states in emergency situations*: In emergency situations, it is important to recognize and address the emotional states of individuals, as fear and stress can greatly impact their behavior and decision-making. Providing reassuring messages can help reduce uncertainties and prevent inappropriate actions, aiding in the development of effective communication strategies that promote informed and appropriate decisions during crises.
f. *Anticipate and address potential effects*: It is important to anticipate the potential negative effects of misinformation, such as inappropriate behaviors, erosion of trust, and the spread of hate speech. Measures to counteract these effects include public awareness campaigns and targeted interventions.
g. *Established network for feedback and response*: Maintaining a well-established network involving relevant actors is crucial to effectively prevent and combat misinformation. Additionally, understanding the dynamics of information networks allows for the implementation of strategies where they can have the most positive impact.
h. *Adaptation to contemporary information systems and technologies*: To address challenges posed by contemporary information systems, such as information overload and the transient nature of information, strategies should be adapted accordingly. Additionally, emerging technologies like AI tools can be utilized not only to combat misinformation and prevent its dissemination but also with caution regarding their potential negative impacts.

To effectively reduce the negative effects of misinformation on risk management, it is essential to implement a range of strategies, from building trust in credible sources to adapting communication for different audiences and platforms. The recommendations in this session emphasize a proactive approach, addressing both the emotional and cognitive aspects of how individuals receive and act on misinformation in crisis situations.

In the next session, we will explore how prebunking and debunking strategies can further strengthen these efforts by countering misinformation before and after it spreads.

6 STEP 5: Prebunk and debunk misinformation

In recent years, advancements in artificial intelligence (AI) tools have provided sophisticated and relatively fast means to fight against earthquake misinformation (Vicari and Komendatova 2023; Komendantova et al. 2021a). These tools could significantly mitigate the adverse effects of misinformation on disaster management, relief efforts, and mitigation policies. Various AI tools employ distinct strategies to tackle misinformation; some are more efficient and effective in certain contexts, while others may be more suitable from alternative perspectives.

AI tools offer powerful solutions for addressing misinformation. Presently, there are various types of AI tools available, including:

a. *Text analysis tools*: This technology enables the analysis and understanding of text, facilitating the identification of misleading claims and detection of language patterns associated with misinformation.
b. *Sentiment analysis tools*: These tools examine the sentiment and tone of content, aiding in the identification of biased or misleading information.
c. *Machine learning algorithms*: These algorithms identify patterns and anomalies in large datasets, helping to detect fake news or disinformation.
d. *Prebunking bots:* These bots counter the spread of misinformation by quickly disseminating accurate information (e.g. the EMSC-developed @LastQuake Twitter bot (Bossu et al. 2023; Fallou et al. 2024)).
e. *Fact-checking bots*: These bots swiftly assess claims and compare them to established facts, enabling real-time debunking of false information.

f. *Content verification tools*: These tools verify the authenticity of images, videos, and audio recordings, making it more difficult for false information to spread.
g. *Source verification tools*: These tools evaluate the credibility and reliability of sources by analyzing their online footprint, history, and associations.
h. *Fact-checking models*: Specifically designed for fact-checking, these models specialize in debunking misinformation within their domain of expertise.

Vicari and Komendatova (2023) conducted a systematic meta-analysis on AI tools for managing misinformation on social media during various hazards and disasters. Their analysis highlighted a significant underrepresentation of social sciences and humanities research, with most studies focusing on COVID-19 and misinformation detection. Limited international funding further restricts the field's development. These findings suggest the necessity of a balanced approach between algorithmic solutions and user autonomy and leveraging pandemic-related research to advance tools for other risks.

In addition to prebunking efforts, manual debunking strategies play a critical role in addressing the nuanced dynamics of earthquake misinformation on social media. Through targeted interventions such as manual tweets and responses to false narratives, seismological institutes like the EMSC can counter misinformation in real-time. In addition, by engaging directly with users and providing empathetic responses that take greater account of users' cultural and emotional needs, manual tweets complement the automated prebunking system and foster trust and credibility among social media users.

6.1 How people perceive AI tools

Together, these AI tools enhance our ability to combat misinformation across various platforms and contexts. However, there are different factors which affect the usefulness and usability of the tools. These factors highlight the complexity and dynamic nature of misinformation, which can vary daily and manifest through different forms and channels. The issue of misinformation is continually evolving, with strategies becoming more sophisticated over time, making it difficult for AI systems to keep pace. Detecting misinformation often requires an understanding of context, cultural nuances, and the intent behind the content (Komendantova et al. 2023). Although AI can identify patterns, it may struggle with the subtleties of human language (Erokhin and Komendantova 2023). Developing AI capable of handling such complex tasks is currently a significant challenge, requiring substantial financial investment to create, train, and maintain effective AI technologies. Many businesses, especially smaller ones, lack the resources to invest in this technology. Another challenge is the level of trust in AI tools. The successful deployment of AI tools depends on user confidence in their objectivity and accuracy. Many individuals and organizations remain skeptical about AI's ability to accurately detect false information. For AI tools to gain wider acceptance, they must demonstrate reliability and fairness.

Given the complexity of the problem and the issue of trust, it is crucial to understand the preferences of various stakeholder groups regarding AI tools for combating misinformation. Preferences refer to an individual's choices or inclinations about what they like, desire, or prioritize (Komendantova et al. 2021b). These choices are shaped by personal experiences, cultural influences, societal norms, and individual values, impacting lifestyle, decision-making, and overall well-being. Studying preferences is vital for several reasons: it aids in decision-making, helps identify compromise solutions among available options, and promotes efficient resource use. Decision-makers need to understand stakeholder preferences to make informed, data-driven decisions, which are essential for developing effective programs, policies, and strategies. This understanding facilitates effective resource allocation, directing resources toward areas with the most significant impact or urgent needs based on stakeholder values. Since different stakeholder preferences often lead to conflicting interests, researching these preferences can help find common ground and resolve conflicts by identifying agreement points or addressing concerns. Studying preferences on AI tool usage is important due to the diverse and heterogeneous nature of the users involved. Currently, AI tools are used by fact-checking organizations, social media platforms, government agencies, tech companies, media outlets, academia, and NGOs. Preferences also vary widely depending on factors such as location, industry, and the specific goals of the users, making it essential to understand these nuances to effectively address misinformation.

Based on the survey results, conducted by EMSC in cooperation with IIASA, slightly under 40% of participants reported encountering misinformation about earthquakes. The most cited types of misinformation include predictions of earthquakes, claims that earthquakes are caused by the HAARP (High-Frequency Active Auroral Research Program) or actions by the United States government, fabricated videos of damage, incorrect earthquake magnitudes, and false casualty numbers. The research findings indicate that most participants believed they had not encountered earthquake misinformation. However, the proportion of those who reported encountering misinformation is nearly equal to those who did not. The most reported types of misinformation include "earthquake predictions", "causes of earthquakes", "fake videos of damage", "incorrect earthquake magnitudes", and "false casualty numbers."

The EMSC X channel (EMSC 2024) shares four types of manually deployed information: “messages to correct false information and misinformation”, “compassion messages in the event of destructive earthquakes”, “responses to specific questions regarding destructive earthquakes”, and “communication about research projects”. Among these, correcting false information and misinformation is deemed the most crucial. The key aspect of messages aimed at preventing the spread of earthquake misinformation is the “content of the message”. Following this are the “timing” of the message, the “level of language” used, “who is sharing the information”, and the “format of the message” (such as visual elements or statistics). Users rated “usefulness” and “trustworthiness” as the key factors in this assessment. Therefore, the top priority for manually crafted posts should be delivering practical and reliable information to address and correct false information and misinformation.

By integrating AI-driven prebunking and debunking strategies, institutions, practitioners and scientists are not only countering misinformation in real-time but also fostering public trust and scientific literacy. These efforts form the foundation for assessing how such measures influence broader communication patterns, as explored in the next session. Evaluating the impact of these interventions will provide valuable insights into their effectiveness and the potential shifts they create in public discourse during disaster situations.

Table 4 SWOT analysis template with guiding questions (Harrison 2024)

Strengths	Weaknesses
What do we do well? What do our target say we do well? What is our unique offer proposition? Do we have strong brand awareness/customer loyalty? Influencer relationships? What skills do we have that our others don’t?	Where can we improve? What do our targets frequently complain about? Which objections are hardest to overcome? Do we have any limitations in delivering? Are our resources and equipment outdated or limited? Are we suffering from skills, or training deficiencies?
Opportunities	Threats
Is there an untapped pain point? Are there potential new sources of support? Are social or political trends that could benefit us? Are any technologies that could benefit us?	Social or political trends that could work against us? Any new technology that could work against us?

7 STEP 6: Evaluate the effectiveness of measures

A thorough review of the strategies implemented to combat misinformation is essential for evaluating their effectiveness. By assessing these strategies, we can determine how well they address the challenges posed by misinformation during disasters. This review helps identify gaps and strengths in current frameworks, guiding improvements to enhance resilience against false information.

7.1 SWOT analysis: Strengths, weaknesses, opportunities, and threats

A SWOT analysis (Teoli et al. 2024) is a strategic planning tool used to evaluate the Strengths, Weaknesses, Opportunities, and Threats (SWOT) associated with a particular situation or decision. When applied to combating misinformation, it involves identifying the strengths and weaknesses of current approaches, such as communication strategies or prebunk and debunk initiatives, in addressing misinformation. Additionally, it explores potential opportunities, such as leveraging new technologies or partnerships, to enhance effectiveness. Moreover, it considers threats posed by factors like the rapid spread of misinformation on social media or public distrust in authoritative sources. By systematically assessing these factors, a SWOT analysis can inform the development of targeted interventions and strategies to mitigate the impact of misinformation during disasters or crises.

An illustrative example comes from Çevik et al.’s study (2024), which applied SWOT analysis to evaluate and improve communication strategies addressing misinformation around HPV vaccination, highlighting key internal and external factors that shape vaccine acceptance among European family doctors and their young patients.

To support such analyses, the example questions presented in Table 4 can help systematically identify relevant strengths, weaknesses, opportunities, and threats, providing a structured foundation for developing more effective communication measures.

7.2 Evaluate how the implementation of measures affects the dynamics of information dissemination and reception during disasters

The analysis techniques presented under Step 2, focusing on misinformation patterns, and Step 3, focusing on misinformation effects, can be used to verify the effectiveness of strategies to combat misinformation and identify opportunities for improvement. Intervention studies can also play a critical role here, testing the effectiveness of various methods such as fact-checking and public information campaigns. Various examples exist in the literature. For instance, Badinathran (2021) conducted a field experiment in India to evaluate a pedagogical intervention's impact on identifying misinformation during the 2019 elections. Yousuf et al. (2021) tested debunking methods to reduce COVID-19 vaccine misinformation in a two-arm randomized blinded parallel study. Ali and Qazi (2023) used a randomized experiment to assess educational interventions to counter misinformation in urban Pakistan among low digital literacy populations.

7.3 Monitor changes in communication patterns and adapt strategies accordingly

Continuous monitoring of communication patterns and their impact is essential to adapt strategies effectively. By observing changes in how information is shared and received, we can identify trends and shifts resulting from the implemented measures. This ongoing assessment helps in promptly addressing any emerging issues and refining strategies to ensure they remain effective. Adapting communication tactics based on real-time data and a dynamic approach ensures that efforts to combat misinformation are responsive and relevant to current conditions.

The continuous monitoring of communication patterns allows us to identify the strengths and weaknesses of current strategies and make necessary adjustments. This dynamic approach ensures that interventions remain effective in real-time scenarios. Moving to the next session, we will explore the ethical considerations that arise when implementing these strategies, especially issues related to transparency, public safety, and freedom of expression in the context of evolving misinformation landscapes. Understanding these ethical challenges will help shape more responsible and inclusive disaster strategies to deal with misinformation.

8 STEP 7: Ethical recommendations and challenges

Effective hazard and risk communication to the public necessitate ethical considerations, such as accessibility, comprehension, and relevance of information. Institutions must ensure that information is readily available, understandable, and reaches all stakeholders, including vulnerable groups. Ethical dilemmas arise when deciding what information to disclose and on what grounds, as well as when balancing scientific evidence with societal values in democratic decision-making processes. While governments have the right to control information dissemination for public health protection, interventions should be limited and monitored, with media acting as independent supervisors. Social media platforms, as significant sources of misinformation, face ethical questions regarding interventions and fostering open debates while protecting users from harm. Ethical discussions are necessary to balance freedom of expression with protecting society from the adverse effects of false information.

8.1 Strategies to monitor and combat misinformation aligned with international human rights law

Strategies to monitor and combat misinformation should align with International Human Rights Law (United Nations 2022). Ensuring freedom of expression is paramount; censorship should be minimized and, when necessary, be transparent, lawful, and proportionate. Encouraging exposure to a variety of opinions and providing context with well-researched arguments can effectively uphold the right to access information. It is also crucial to combat hate speech and discrimination to maintain a respectful and informative public discourse.

8.2 Complying with the digital services act

The Digital Services Act (DSA) (European Commission 2025) is a regulatory framework that intersects with human rights considerations in tackling misinformation. Enacted by the European Union, the DSA establishes clear responsibilities for online platforms to mitigate risks, enhance transparency, and ensure accountability in digital environments. It mandates proactive measures against illegal content while safeguarding users' fundamental rights, including freedom of expression and privacy. Large platforms and search engines must conduct risk assessments and implement due diligence obligations to prevent the spread of harmful misinformation. By enforcing transparency in content moderation and algorithmic decision-making, the DSA aims to balance the fight against disinformation with the protection of democratic freedoms.

8.3 Adhering to the general data protection regulation

Adhering to the General Data Protection Regulation (GDPR) (European Parliament and European Council 2016) is essential in this context. The GDPR, which came into effect in May 2018, is a comprehensive regulation that provides robust protections for personal data of European Union citizens. It enforces strict privacy rules and grants individual rights such as the right to access their data, the right to rectify inaccuracies, and the right to erasure, often referred to as the "right to be forgotten". When analyzing misinformation, preference should be given to aggregated data, and users' identities should be anonymized to protect privacy. Organizations must ensure that data processing activities are lawful, fair, and transparent, maintaining data integrity and confidentiality.

8.4 Implementing the EU AI Act

The implementation of the EU Artificial Intelligence (AI) Act (European Commission 2024) also plays a critical role. This act, proposed in April 2021, aims to create a harmonized regulatory framework for AI across the EU. It categorizes AI systems based on their risk levels, ranging from minimal risk to unacceptable risk. For AI systems used to counter misinformation, which often fall into the high-risk category, the act mandates stringent requirements for transparency, accuracy, and accountability. Ensuring transparency of training data is essential to guarantee that AI systems are ethical and trustworthy. The AI Act also requires that high-risk AI systems undergo conformity assessments before being deployed, ensuring they meet the required standards for safety and ethical considerations. The International Human Rights Law (United Nations 2023), the General Data Protection Regulation (GDPR) (European Parliament and European Council 2016) and the EU Artificial Intelligence Act (European Commission 2024) and their interactions are outlined in Fig. 3.

8.5 Ethics for natural language processing tools

Natural Language Processing (NLP) is a branch of AI focused on the interaction between computers and human language. It includes several key subsets: Natural Language Understanding (NLU), which allows machines to comprehend meaning and context; Natural Language Generation (NLG), enabling machines to produce human-like text; Machine Translation, which automatically converts text between languages; and Large Language Models (LLMs), advanced models that generate and understand complex language patterns. Together, these technologies enable machines to process, interpret, and generate language, driving innovations in areas such as text analysis, conversational agents, and automated content creation.

Vicari and Komendatova (2024, in preparation) conducted a meta-analysis of studies published in the Web of Science, focusing on "Ethics in the field of Natural Language Processing". They identified 465 relevant papers, excluding duplicates and reviews, and ultimately selected 208 publications as significant for their research. Each study was analyzed for its objective, ethical principles, sponsor location, year of publication, research area, and technology type. The analysis revealed a peak in publications in 2023, with 133 papers, predominantly on LLMs (77 papers). Most research occurred in the sector of healthcare (60 studies), with the USA as a key funder. The primary research focus in the corpus of studies was on identifying ethical challenges (113 papers) and defining ethical standards (84

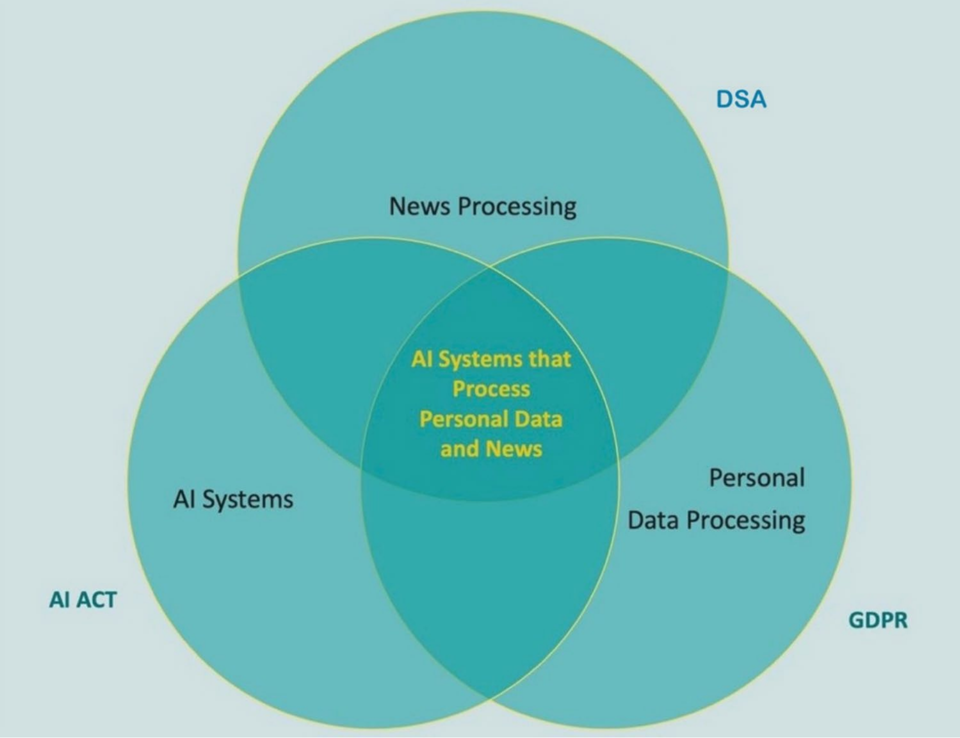


Fig. 3 Strategies to monitor and combat misinformation should align with Digital Services Act (European Commission 2025), the General Data Protection Regulation (GDPR) (European Parliament and European Council 2016) and the EU AI Act (European Commission 2024)

papers), while fewer studies addressed implementing ethical standards (12) or designing ethical tools (30). The most frequently mentioned ethical principles were accuracy and misinformation prevention (80 papers), followed closely by privacy (79 papers), with other frequent principles including bias mitigation (61 papers), transparency (57 papers), safety (63 papers), and intellectual property (55 papers).

The results suggest a growing emphasis on ethics in NLP, particularly in relation to LLMs, as evidenced by the significant increase in publications in 2023 related to this technology. The concentration of research in the healthcare sector indicates a heightened awareness of ethical implications where NLP impacts sensitive areas like patient care. By leveraging ethical frameworks and tools from healthcare, where accuracy, privacy, bias prevention, and safety are critical, researchers can adapt these strategies for misinformation detection, prevention, and mitigation in hazards and disasters. This approach would help maintain AI tools as reliable and ethically sound in high-stakes conditions.

The dominant focus on identifying ethical challenges and defining standards highlights the field's ongoing struggle to navigate complex ethical issues. However, the relatively few studies on implementing these standards or designing ethical tools suggest a gap between theoretical discussions and practical applications, pointing to a need for more actionable research. Additionally, the prominent role of the United States as a funder emphasizes the need for more diverse, global perspectives in this research area.

The increasing focus on ethics in Natural Language Processing, particularly in high-stake areas like healthcare, underscores the need for actionable frameworks that can also be applied to managing disaster-related misinformation. While significant progress has been made in identifying ethical challenges, more emphasis on the practical implementation is required. As we move to the next session, we will focus on the managerial and operational aspects of implementing these strategies, examining how policies and guidelines can be structured to ensure that misinformation management is both effective and aligned with public safety during emergencies.

9 STEP 8: Managerial aspects: policy and operational implementation guidelines

Misinformation can significantly impact emergency management actions before, during, and after an event. It may cause a cascading effect, as misinformation can lead to undesirable public behavior, which in turn can exacerbate the situation or create new emergencies. For instance, during evacuations, if the public receives misleading information from rumors about an evacuation route, this misinformation could lead to dangerous, even deadly outcomes. Misinformation poses a substantial threat to the effectiveness of emergency management actions, particularly in interactions with the public. Furthermore, misinformation can result in panic behavior or confusion, especially when it contradicts official instructions or reports false events. For example, during the early stages of the COVID-19 pandemic, widespread misinformation about food and supply shortages led to panic buying in many countries.

9.1 Erosion of public trust and the necessity of effective communication

Misinformation can erode public trust in emergency management preparations and diminish confidence in the actions taken by authorities. When the public loses trust, they are less likely to follow official guidelines and instructions, which can lead to chaotic and uncoordinated responses. Moreover, emergency managers might implement inappropriate policies based on incorrect data, leading to resource misallocation and delayed responses.

Coping with misinformation requires close interaction with the public and their collaboration. Establishing clear, consistent, and transparent communication channels to disseminate accurate information is crucial for building trust among citizens and promoting awareness of reliable information sources during routine times and emergencies. It is essential for citizens to trust and follow instructions and information provided by well-known and trusted sources, a relationship that should be established well before an emergency.

9.2 Real-time monitoring and public education

Emergency managers must recognize that the spread of misinformation is inevitable, even in short and local events. Real-time monitoring is vital to address misinformation promptly, allowing emergency managers to provide relevant and reliable information quickly. Utilizing social media and other platforms to monitor and counteract misinformation swiftly is essential. Educating the public on detecting misinformation and developing critical thinking about behavior during emergencies is also crucial. Public awareness campaigns can teach the public how to recognize and report misinformation and encourage fact-checking to verify and debunk false information.

9.3 Proactive messaging and legal measures

Fighting misinformation involves proactive messaging and information management. By providing regular updates and accurate information, authorities can address potential misinformation preemptively, minimizing negative and undesirable outcomes. A centralized information hub, regularly updated during emergencies, can serve as a reliable source for the public to verify facts. Partnerships with both local and national media can help diminish misinformation and spread correct information and instructions. Working closely with reputable media outlets ensures that accurate information is broadcast. Lastly, misinformation should be addressed with legal sanctions. Educating the public on detecting trusted information and ignoring misinformation can be reinforced by legal measures, aiding in the fight against the spread of false information.

9.4 Eight management principles to effectively deal with hazard and disaster misinformation management

We defined eight guiding principles based on insights from a communication guide developed by the Euro-Mediterranean Seismological Centre (EMSC) and the Swiss Seismological Service at ETH Zurich (Dallo et al. 2022a) to assist institutions, scientists, and practitioners in communicating earthquake information to the public and combating misinformation. We have adapted these guidelines for broader application. These principles, depicted in Fig. 4, can be implemented in the context of any hazard or disaster (whether anthropogenic or natural) to enhance communication efforts and effectively counter misinformation.

9.4.1 Assess and understand the audience

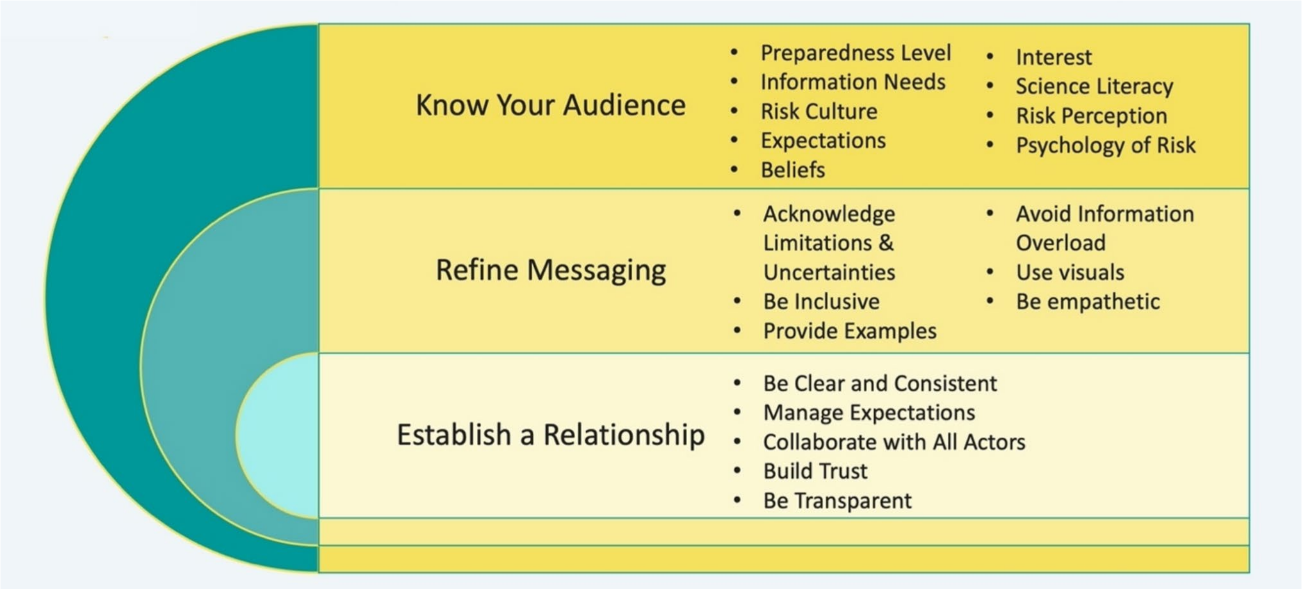
Effective risk communication requires understanding audience beliefs, cultural contexts, and exposure to misinformation to tailor strategies that build trust. For instance, rural communities with older residents and misconceptions about earthquake risks may respond better to trusted local channels than to social media. Equally vital is grasping how misinformation originates and spreads. By addressing misconceptions clearly and empathetically, communicators can reduce misinformation's impact and support informed, resilient communities (Vraga and Bode 2017; Southwell et al. 2017).

9.4.2 Build and maintain trust

Building relationships with communities and stakeholders before a crisis is fundamental to establishing trust and ensuring effective communication during emergencies. Ongoing engagement through regular interaction and participation in community events fosters credibility and facilitates dialogue that respects both scientific knowledge and local perspectives. Additionally, clearly managing expectations by communicating what information will be available, when it will be shared, and through which channels can reduce uncertainty and limit the spread of misinformation during crises (Steelman and McCaffrey 2013; Covello 2003).

9.4.3 Use effective communication techniques

Refining risk communication through clear messaging and timely delivery is essential for enhancing public



Know Your Audience

- Preparedness Level
- Information Needs
- Risk Culture
- Expectations
- Beliefs
- Interest
- Science Literacy
- Risk Perception
- Psychology of Risk

Refine Messaging

- Acknowledge Limitations & Uncertainties
- Be Inclusive
- Provide Examples
- Avoid Information Overload
- Use visuals
- Be empathetic

Establish a Relationship

- Be Clear and Consistent
- Manage Expectations
- Collaborate with All Actors
- Build Trust
- Be Transparent

Fig. 4 Originally created to help institutions, scientists, and practitioners communicate earthquake information and combat misinformation, we have adapted these principles for broader application (Dallo et al. 2022b)

understanding and countering misinformation during disasters. The use of visuals and concrete examples can significantly improve message comprehension, especially when rapid understanding is critical. Providing clear, relevant, and accessible information to diverse audiences helps sustain attention and minimize confusion. Moreover, aligning communication efforts with the phases of a disaster—before, during, and after—ensures that emerging misinformation is addressed proactively and continuously throughout the crisis cycle (Reynolds and Seeger 2005; Lachlan et al. 2014).

9.4.4 Address misinformation proactively

Combating misinformation requires identifying prevalent myths and providing evidence-based explanations tailored to the audience's level of scientific literacy. Continuous and sustained engagement by trusted institutions is key to effectively countering falsehoods and maintaining public trust. Since misinformation can emerge at any point in the disaster cycle, ongoing monitoring and responsive communication are essential components of a comprehensive strategy (Linden et al. 2017; Lewandowsky et al. 2012).

9.4.5 Engage continuously and monitor diligently

Sustained and effective risk communication relies on regular engagement with target audiences, continuous testing of communication materials, and the flexibility to adapt strategies based on feedback and emerging issues. Permanent monitoring is crucial to quickly identify misconceptions and respond proactively. Equally important is allocating sufficient resources and ensuring involvement of trained, transdisciplinary personnel who can engage diverse audiences with credibility and competence, thereby strengthening the overall impact of communication initiatives (Veil et al. 2011; Sellnow and Seeger 2020).

9.4.6 Prepare and train for crisis communication

Integrating communication strategies into emergency plans is essential to ensure effective outreach when crises unfold and resources are strained. Anticipating communication challenges and proactively embedding them into preparedness efforts helps mitigate the spread of misinformation and maintain public trust. Even under difficult circumstances, sustained communication is crucial, as silence can foster uncertainty and erode credibility. Individual and institutional efforts to communicate clearly and consistently can significantly shape public response and resilience (Reynolds and Seeger 2005; Seeger 2006).

9.4.7 Acknowledge the complexity of science

Transparent communication about the uncertainties and limitations of scientific knowledge is essential for building trust and fostering a dialogue of equals with the public. Acknowledging what is known and unknown helps manage expectations, reduces the risk of overconfidence, and addresses misbeliefs respectfully. Such openness enhances credibility and encourages informed decision-making, especially in complex or evolving risk situations (Fischhoff 2012; Gustafson and Rice 2019).

9.4.8 Sustain continuous engagement

A long-term, consistent commitment by institutions is crucial for effectively combating misinformation and maintaining public trust. Continuous engagement ensures that communication remains responsive to evolving audience needs, reinforces credibility, and supports the sustained dissemination of accurate information. Such enduring involvement allows institutions to build relationships, adapt to changing contexts, and counter misinformation more effectively over time (Southwell et al. 2017).

In conclusion, applying these eight management principles in hazard and disaster communication will not only strengthen the ability of institutions, practitioners and scientists to counter misinformation but it will also foster a deeper connection with their audience. By tailoring the communication strategy to the unique needs and perceptions of different communities, and maintaining a proactive, transparent, and consistent communication strategy, it is possible to effectively build trust and resilience in times of crisis. Ultimately, these principles serve as a practical roadmap for turning communication challenges into opportunities for creating more informed and prepared communities.

10 Conclusions and perspectives

The methodology presented in this technical note offers a comprehensive toolbox for assessing, preventing and mitigating misinformation risks and impacts in disaster risk management (DRM). By systematically addressing the multifaceted challenges posed by misinformation on natural and anthropogenic hazards and disasters, this framework aids researchers, institutions, policy makers, decision makers, and practitioners in developing robust strategies to enhance public trust and response efficacy.

The eight-step approach integrates diverse analytical techniques and tools, from defining the communication context to employing advanced AI tools for prebunking and debunking misinformation. The methodology emphasizes the importance of a tailored, context-specific application, recognizing that not all tools and steps need to be exhaustively implemented in every situation. This flexibility ensures that strategies remain relevant and effective across different disaster scenarios.

Key findings highlight the critical role of understanding communication patterns, identifying misinformation sources and impacts, and implementing ethical, legal, and managerial measures. The inclusion of stakeholder preferences and the need for continuous monitoring and adaptation of strategies underscore the dynamic nature of misinformation and the necessity for ongoing vigilance.

Looking ahead, the methodology outlined in this technical note presents several avenues for further research and practical applications. In this study, we have presented different disaster and misinformation case studies to illustrate each methodological step, as shown in Table 5; some case studies involved two different steps. In the future, it would be interesting to test multiple methodological steps together on a single case study or a selection of case studies corresponding to different disaster contexts. Applying the full framework to a single case study or through pilot validations in varied contexts (e.g. different hazard contexts) would allow for assessing its practical feasibility, adaptability, and impact. Importantly, a comprehensive evaluation of the approach would require several years of implementation across the disaster risk management cycle, from the prevention and preparedness phase, through early warning and emergency response, to recovery and long-term impact assessment (UNDDR 2015; Coppola 2021). Such longitudinal application would enable systematic testing of the framework's effectiveness in supporting risk communication and misinformation management over time and across different stages of disaster risk reduction. Further research should also explore the adaptation of the framework in diverse sociopolitical and cultural backgrounds, such as non-Western contexts. In particular, low-resource settings may face structural challenges such as limited access to digital infrastructures, reduced institutional capacity for coordinated communication, and greater reliance on informal channels of information exchange. Authoritarian environments, by contrast, may

Table 5 Summary table presenting each methodological step, the associated tools, and the related case studies

Eight methodological steps	Tools and corresponding case studies
STEP 1: Define the communication context	• Pestel analysis (Kung 2023) • Berlo's communication model (Dallo 2022)
STEP 2: Identify current misinformation patterns	• Natural language processing methods (Elroy and Yosipof 2022; Elroy et al. 2023; Dallo et al. 2023) • Machine Learning (Elroy and Yosipof 2023) • Natural language processing and Machine learning Model Performance (Elroy and Yosipof 2022, 2023; Elroy et al. 2023; Dallo et al. 2023; Vicari et al. 2024) • Time Series (Erokhin et al. 2022; Elroy et al. 2023; Dallo et al. 2023) • Content analysis (Elroy et al. 2023; Dallo et al. 2023) • User engagement (Dallo et al. 2023; Vicari et al. 2024) • Sentiment analysis (Vicari et al. 2024)
STEP 3: Assess misinformation impact on risk perceptions and risk management	• Conceptual and theoretical framework (Pundir et al. 2021; Griffin et al. 2004; Hansson et al. 2020) • Surveys and questionnaires (Griffin et al. 2004) • Focus Groups (Hertzum et al. 2002) • Interviews (Das and Ahmed 2022) • Delphi Method (Flostrand et al. 2019)
STEP 4: Implement measures to mitigate negative effects	Recommendations for preventing and combating the spread of misinformation across the entire communication chain (Dallo et al. 2023; Elroy and Yosipof 2022, 2023; Dallo et al. 2023; Vicari et al. 2024; Gugg 2024; Rapaport et al. 2024)
STEP 5: Prebunk and debunk misinformation	•AI tools to prebunk and debunk misinformation related to disasters (Vicari and Komendatova 2023; Komendantova et al. 2021a) • Users' preferences (Komendantova et al. 2023, 2021b; Erokhin and Komendantova 2023)
STEP 6: Evaluate the effectiveness of measures	• SWOT analysis (Çevik et al. 2024) • Evaluate the impact of measures on information dynamics (Badrinathan 2021)
STEP 7: Ethical recommendations and challenges	Current regulation and challenges in the field of Natural Language Processing (Vicari and Komendatova 2024, in preparation)
STEP 8: Managerial aspects: policy and operational implementation guidelines	Management principles (Dallo et al. 2022a, 2022b)

involve restricted information flows, state-controlled narratives, and heightened risks for civil society actors. These contextual differences can substantially shape how misinformation emerges, spreads, and is countered. Therefore, systematic cross-cultural validation of the framework is critical to ensure its robustness, inclusivity, and applicability across a wide range of governance and communication environments. Additionally, we could explore how the adaptability of this framework is influenced by different cultural backgrounds, particularly in non-Western contexts, offering valuable insights for broadening its global application and ensuring the tools' effectiveness across diverse social and cultural settings. Moreover, the integration of real-time data analytics and AI advancements holds promise for even more rapid and accurate identification and mitigation of misinformation.

The co-creation approach involving the public highlights the potential for greater community engagement and education in DRM. Strengthening public awareness and critical thinking skills through targeted educational campaigns can further enhance resilience against misinformation. Additionally, exploring the interplay between cognitive biases and misinformation can provide deeper insights into tailoring interventions to effectively counteract these biases.

Ethical considerations will remain paramount as technological tools evolve. Ensuring compliance with international human rights laws and regulations, such as the, DSA, GDPR and the EU AI Act, is essential for maintaining public trust and protecting individual rights. Implementing ethical standards and designing ethical tools will be crucial in navigating these ethical challenges through actionable research.

By implementing real-time monitoring, proactive messaging, and building public trust through clear and consistent communication, emergency managers can better navigate the challenges misinformation poses. The operational guidelines outlined in the last session provide a practical framework for institutions to maintain public trust and enhance their communication strategies, ensuring that misinformation does not undermine emergency efforts.

In conclusion, the toolbox presented in this technical note provides a robust foundation for tackling misinformation in disaster risk management. By continuing to refine and adapt these methodologies, stakeholders can build more resilient communities capable of effectively responding to the complex challenges of misinformation in disaster contexts.

11 Declaration of generative AI in scientific writing

During the preparation of this work, the author(s) used ChatGPT 3.5 in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00146-025-02680-5.

Author contributions **R.V.** and **N.K.** conceptualized the review. **R.V.**, **N.K.**, **O.E.**, **I.D.**, **C.R.**, **C.D.** and **A.Y.** contributed to develop the described methodology. **R.V.**, **N.K.**, **O.E.**, **I.D.**, and **A.Y.** curated the data. **R.V.** prepared Figs. 1–4 and Tables 1–2 and 4. **I.D.** prepared Table 3. **R.V.**, **N.K.**, **O.E.** and **A.Y.** drafted the main manuscript text. **I.D.**, **C.R.** and **C.D.** contributed to parts of the draft manuscript. **R.V.** edited the manuscript text. **N.K.** and **A.Y.** supervised and validated the work. All authors reviewed the manuscript.

Funding Open access funding provided by International Institute for Applied Systems Analysis (IIASA). The author(s) disclosed receipt of the following financial support for the research, authorship and publication of this article: [details omitted for double-anonymized peer review].

Data availability No datasets were used or generated in the preparation of this technical note.

Declarations

Conflict of interests The authors declare no competing interests.

References

Aguilar FJ (1967) Scanning the business environment. Macmillan, Johannesburg

Ali A, Qazi IA (2023) Countering misinformation on social media through educational interventions: evidence from a randomized experiment in Pakistan. J Dev Econ 163:103108. https://doi.org/10.1016/j.jdeveco.2023.103108

Badrinathan S (2021) Educative interventions to combat misinformation: evidence from a field experiment in India. Am Polit Sci Rev 115(4):1325–1341. https://doi.org/10.1017/S0003055421000459

Bossu R, Corradini M, Cheny J-M, Fallou L (2023) A social bot in support of crisis communication: 10-years of @LastQuake experience on Twitter. Front Commun. https://doi.org/10.3389/fcomm.2023.992654

Çevik HS, Peker AGC, Görpelioğlu S, Vinker S, Ungan M (2024) How to overcome information and communication barriers in Human Papillomavirus vaccination? A SWOT analysis based on the opinions of European family doctors in contact with young people and their parents. Eur J Gen Pract. https://doi.org/10.1080/13814788.2024.2393858

Cision (2023) Europresse. Available: https://www.europresse.com. Accessed 11 Nov 2023 [Online].

Coppola DP (2021) Introduction to international disaster management, 4th edn. Butterworth-Heinemann.

Covello VT (2003) Best practices in public health risk and crisis communication. J Health Commun 8(1):5–8. https://doi.org/10.1080/713851971

Dallo I (2022) Understanding the communication of event-related earthquake information in a multi-hazard context to improve society's resilience, ETH Zurich, Zurich. Available: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/535657/1/Doctoralthesis_DalloIrina_Multi-hazardcommunication.pdf. Accessed 02 June 2025 [Online].

Dallo I, Elroy O, Fallou L, Komendantova N, Yosipof A (2023) Dynamics and characteristics of misinformation related to earthquake predictions on Twitter. Sci Rep 13(1):13391. https://doi.org/10.1038/s41598-023-40399-9

Dallo I, Corradini M, Laure F, Michèle M (2022) How to fight misinformation about earthquakes? A communication guide, Zurich. Available: https://www.research-collection.ethz.ch/handle/20.500.11850/530319. Accessed 01 Sep 2024. [Online].

Dallo I, Corradini M, Fallou L, Marti M (2022) How to fight misinformation about earthquakes? A communication guide, Swiss Seismological Service at ETH Zurich. Available: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/530319/CommunicationGuide_FightingEarthquakeMisinformation_Dallo_Corradini_Fallou_Marti.pdf?sequence=3&isAllowed=y Accessed 13 Nov 2022. [Online].

Dallo I et al. (2023) Impact of misinformation on social media on risk perception in a multi-risk environment. Available: https://pure.iiasa.ac.at/id/eprint/19767/. Accessed 01 Sep 2024. [Online].

Das R, Ahmed W (2022) Rethinking fake news: disinformation and ideology during the time of COVID-19 global pandemic. IIM Kozhikode Soc Manage Rev 11(1):146–159. https://doi.org/10.1177/22779752211027382

DiFonzo N, Bordia P (2007) Rumor psychology: social and organizational approaches. Washington: American Psychological Association. https://doi.org/10.1037/11503-000.

Dubetcky O (2024) Re-imagine the business strategy: PEST/EL Analysis, Medium. Available: https://oleg-dubetcky.medium.com/re-imagine-the-business-strategy-pest-el-analysis-2a77e614f4d6. Accessed 03 Sep 2024. [Online].

Elroy O, Yosipof A (2022) Analysis of COVID-19 5G conspiracy theory tweets using sentence BERT embedding, pp 186–196. https://doi.org/10.1007/978-3-031-15931-2_16.

Elroy O, Yosipof A (2023) Semi-supervised learning classifier for misinformation related to earthquakes prediction on social media, pp 256–267. https://doi.org/10.1007/978-3-031-44207-0_22.

Elroy O, Erokhin D, Komendantova N, Yosipof A (2023) Mining the discussion of monkeypox misinformation on twitter using RoBERTa, pp 429–438. https://doi.org/10.1007/978-3-031-34111-3_36.

EMSC (2024) EMSC X Channel. Available: https://x.com/lastquake?lang=en. Accessed 05 Sep 2024. [Online].

Erokhin D, Komendantova N (2023) The role of bots in spreading conspiracies: case study of discourse about earthquakes on Twitter. Int J Disaster Risk Reduct 92:103740. https://doi.org/10.1016/j.ijdrr.2023.103740

Erokhin D, Yosipof A, Komendantova N (2022) COVID-19 conspiracy theories discussion on Twitter. Social Media + Society 8(4):205630512211260. https://doi.org/10.1177/20563051221126051

European Commission (2024) AI Act. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Accessed 01 Sep 2024. [Online].

European Commission (2025) The Digital Services Act. Available: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en. Accessed 16 Mar 2025. [Online].

European Parliament and European Council (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 04 May 2016. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504. Accessed 01 Sep 2024. [Online].

Fallou L, Bossu R, Cheny J-M (2024) Prebunking earthquake predictions on social media. Front Commun. https://doi.org/10.3389/fcomm.2024.1391480

Fischhoff B (2012) Communicating uncertainty: fulfilling the duty to inform. Issues Sci Technol 28(4):63–70

Flostrand A, Pitt L, Kietzmann J (2019) Fake news and brand management: a Delphi study of impact, vulnerability and mitigation. J Prod Brand Manage 29(2):246–254. https://doi.org/10.1108/JPBM-12-2018-2156

Griffin RJ, Neuwirth K, Dunwoody S, Giese J (2004) Information sufficiency and risk communication. Media Psychol 6(1):23–61. https://doi.org/10.1207/s1532785xmep0601_2

Gugg G (2024) Afar from vesuvius but still at risk. In: Disasters and changes in society and politics, Bristol University Press, 2024, pp 102–117. https://doi.org/10.2307/jj.9692578.11.

Gustafson A, Rice RE (2019) The effects of uncertainty frames in three science communication topics. Sci Commun. https://doi.org/10.1177/1075547019870811

Hansson S et al (2020) Communication-related vulnerability to disasters: a heuristic framework. Int J Disaster Risk Reduct 51:101931. https://doi.org/10.1016/j.ijdrr.2020.101931

Harrison K (2024) Use SWOT, PESTLE and VUCA analysis for communication planning, Cutting edge PR. Available: https://cuttingedgepr.com/articles/use-swot-pestle-vuca-analysis-for-communication-planning. Accessed 03 Sep 2024. [Online].

Hertzum M, Andersen HHK, Andersen V, Hansen CB (2002) Trust in information sources: seeking information from people, documents, and virtual agents. Interact Comput 14(5):575–599. https://doi.org/10.1016/S0953-5438(02)00023-1

Hunt K, Wang B, Zhuang J (2020) Misinformation debunking and cross-platform information sharing through Twitter during Hurricanes Harvey and Irma: a case study on shelters and ID checks. Nat Hazards 103(1):861–883. https://doi.org/10.1007/s11069-020-04016-6

Hutto C, Gilbert E (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Social Media 8(1):216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Ireton C, Posetti J (2018) Journalism, fake news & disinformation: handbook for journalism education and training. UNESCO, 2018. Available: https://unesdoc.unesco.org/ark:/48223/pf0000265552. Accessed 22 Jul 2024. [Online].

Ireton C, Posetti J (2018) Journalism, 'fake news' & disinformation, 2018, United Nations Educational, Scientific and Cultural Organization. Available: https://en.unesco.org/sites/default/files/journalism_fake_news_disinformation_print_friendly_0_0.pdf. Accessed 13 Nov 2022. [Online].

Jensen E, Laurie C (2016) Doing real research : a practical guide to social research, SAGE Publications Ltd. Available: http://digital.casalini.it/9781473944299. Accessed 01 Sep 2024. [Online].

Komendantova N et al (2021a) A value-driven approach to addressing misinformation in social media. Humanit Soc Sci Commun 8(1):33. https://doi.org/10.1057/s41599-020-00702-9

Komendantova N, Erokhin D, Albano T (2023) Misinformation and its impact on contested policy issues: the example of migration discourses. Soc (Basel) 13(7):168. https://doi.org/10.3390/soc13070168

Komendantova N, Ekenberg L, Amann W, Danielson M, Koulolias V (2021) Chapter 10 The adequacy of artificial intelligence tools to combat misinformation, pp 172–198. https://doi.org/10.1007/978-3-030-70370-7_10.

Kung W (2023) Using the PESTEL analysis to determine the effectiveness of new digital media strategies. Adv Econom, Manage Polit Sci 5(1):19–25. https://doi.org/10.54254/2754-1169/5/20220054

Lachlan KA, Spence PR, Lin X, Del Greco M (2014) Screaming into the wind: examining the volume and content of tweets associated with Hurricane Sandy. Commun Stud 65(5):500–518. https://doi.org/10.1080/10510974.2014.956941

Lazer DMJ et al (2018) The science of fake news, Science, 359(6380):1094–1096. https://doi.org/10.1126/science.aao2998.

Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction. Psychol Sci Public Interest 13(3):106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky S, Ecker UKH, Cook J (2017) Beyond misinformation: understanding and coping with the 'post-truth' era. J Appl Res Mem Cogn 6(4):353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Liu T, Xiao X (2021) A framework of AI-based approaches to improving eHealth literacy and combating infodemic. Front Public Health. https://doi.org/10.3389/fpubh.2021.755808

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, In: 2020 IEEE International Conference on Big Data (Big Data), IEEE, Dec 2020, pp 748–757. https://doi.org/10.1109/BigData50022.2020.9377956.

Naeem SB, Boulos MNK (2021) COVID-19 misinformation online and health literacy: a brief overview. Int J Environ Res Public Health 18(15):8091. https://doi.org/10.3390/ijerph18158091

Nandonde FA (2019) A PESTLE analysis of international retailing in the East African Community. Glob Bus Organ Excel 38(4):54–61. https://doi.org/10.1002/joe.21935

Oh O, Agrawal M, Rao HR (2013) Community intelligence and social media services: a rumor theoretic analysis of tweets during social crises. MIS Q 37(2):407–426. https://doi.org/10.25300/MISQ/2013/37.2.05

Paek H-J, Hove T (2019) Effective strategies for responding to rumors about risks: the case of radiation-contaminated food in South Korea. Public Relat Rev 45(3):101762. https://doi.org/10.1016/j.pubrev.2019.02.006

Papakyriakopoulos O, Medina Serrano JC, Hegelich S (2020) The spread of COVID-19 conspiracy theories on social me-dia and the effect of content moderation, Harvard Kennedy School Misinformation Review, 2020, https://doi.org/10.37016/mr-2020-034.

Pennycook G, Rand DG (2019) Fighting misinformation on social media using crowdsourced judgments of news source quality. Proc Natl Acad Sci USA 116(7):2521–2526. https://doi.org/10.1073/pnas.1806781116

Pian W, Chi J, Ma F (2021) The causes, impacts and countermeasures of COVID-19 'Infodemic': a systematic review using narrative synthesis. Inf Process Manag 58(6):102713. https://doi.org/10.1016/j.ipm.2021.102713

Pundir V, Devi EB, Nath V (2021) Arresting fake news sharing on social media: a theory of planned behavior approach. Manag Res Rev 44(8):1108–1138. https://doi.org/10.1108/MRR-05-2020-0286

Rapaport C, Dallo I, Kligman Y, Marti M, Komendantova N, Ashkenazi I (2024) The same but different: a cross-country comparison of national earthquake policies and societal perspectives of seismic risk in Israel and Switzerland. Risk Hazards Crisis Public Policy. https://doi.org/10.1002/rhc3.12316

Reynolds B, Seeger MW (2005) Crisis and emergency risk communication as an integrative model. J Health Commun 10(1):43–55. https://doi.org/10.1080/10810730590904571

Salehinejad S, Jangipour Afshar P, Borhaninejad V (2021) Rumor surveillance methods in outbreaks: a systematic literature review. Health Promot Perspect 11(1):12–19. https://doi.org/10.34172/hpp.2021.03

Seeger MW (2006) Best practices in crisis communication: an expert panel process. J Appl Commun Res 34(3):232–244. https://doi.org/10.1080/00909880600769944

Sellnow TL, Seeger MW (2020) Theorizing crisis communication, 2nd edn. Wiley-Blackwell.

Southwell BG, Thorson E, Sheble L (2017) The persistence and peril of misinformation. Am Sci 105(6):372. https://doi.org/10.1511/2017.105.6.372

Spence PR, Lachlan KA, Edwards A, Edwards C (2016) Tweeting fast matters, but only if I think about it: information updates on social media. Commun Q 64(1):55–71. https://doi.org/10.1080/01463373.2015.1100644

Steelman TA, McCaffrey S (2013) Best practices in risk and crisis communication: implications for natural hazards management. Nat Hazards 65(1):683–705. https://doi.org/10.1007/s11069-012-0386-z

Teoli D, Sanvictores T, An J (2024) SWOT analysis. StatPearls Publishing

Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press.

United Nations (2022) Universal declaration of human rights. Available: https://www.ohchr.org/en/universal-declaration-of-human-rights. Accessed 13 Nov 2022. [Online].

United Nations (2023) UN actions against hate speech, International Human Rights Law. Available: https://www.un.org/en/hate-speech/united-nations-and-hate-speech/international-human-rights-law. Accessed 11 Nov 2023. [Online].

UNDDR (2015) Sendai framework for disaster risk reduction 2015–2030. Available: https://www.undrr.org/quick/11409. Accessed 29 Sep 2025. [Online].

van der Linden S, Leiserowitz A, Rosenthal S, Maibach E (2017) Inoculating the public against misinformation about climate change. Glob Chall. https://doi.org/10.1002/gch2.201600008

Varma R, Verma Y, Vijayvargiya P, Churi PP (2021) A systematic survey on deep learning and machine learning approaches of fake news detection in the pre- and post-COVID-19 pandemic. Int J Intell Comput Cybern 14(4):617–646. https://doi.org/10.1108/IJICC-04-2021-0069

Veil SR, Buehner T, Palenchar MJ (2011) A work-in-process literature review: incorporating social media in risk and crisis communication. J Contingencies Crisis Manage 19(2):110–122. https://doi.org/10.1111/j.1468-5973.2011.00639.x

Vicari R, Komendatova N (2023) Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters. Humanit Soc Sci Commun 10(1):332. https://doi.org/10.1057/s41599-023-01838-0

Vicari R, Elroy O, Komendantova N, Yosipof A (2024) Persistence of misinformation and hate speech over the years: the Manchester Arena bombing. Int J Disaster Risk Reduct 110:104635. https://doi.org/10.1016/j.ijdrr.2024.104635

Vicari R, Komendatova N (2024) Ethics for natural language processing: a systematic meta-analyis (in preparation)

Vraga EK, Bode L (2017) Using expert sources to correct health misinformation in social media. Sci Commun 39(5):621–645. https://doi.org/10.1177/1075547017731776

Wardle C, Derakhshan H (2017) Information disorder: toward an interdisciplinary framework for research and policy making, Strasbourg 2017. Available: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c. Accessed 29 Mar 2024. [Online].

Wisner B, Blaikie P, Cannon T, Davis I (2004) At risk: natural hazards, people's vulnerability and disasters, 2nd edn. Routledge

Yosipof A, Woo G, Komendantova N (2023) Persistence of risk awareness: Manchester arena bombing on 22 May 2017. Int J Disaster Risk Reduct 94:103805. https://doi.org/10.1016/j.ijdrr.2023.103805

Yosipof A, Rapaport C (2023) Report about communication patters. Available: https://www.euproject-core.eu/images/deliverables/CORE-D7.1-Report%20about%20communication%20patters.pdf Accessed 05 Sep 2024. [Online].

Yousuf H et al (2021) A media intervention applying debunking versus non-debunking content to combat vaccine misinformation in elderly in the Netherlands: a digital randomised trial. EClinicalMedicine 35:100881. https://doi.org/10.1016/j.eclinm.2021.100881

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.