********** Accepted Version To appear in Journal of Risk Research**********

A risk-analysis framework for evaluating the impact of information disorders

Magda Osman[1,2]

m.osman@jbs.cam.ac.uk

1. Judge Business School, University of Cambridge, Trumpington St, Cambridge CB2 1AG.

2. Centre for Decision Research, Leeds Business School, University of Leeds, Leeds, West Yorkshire, LS2 9JT, UK.

Abstract

What kind of risk-based framework could support an evaluation of impacts resulting from information disorders (e.g., misinformation, disinformation, fake news)? To evaluate their consequences, the starting point proposed in this article is to set out criteria to prioritise which type, and which domain to locate them in. To solve the quagmire of defining what is and isn't disordered information, the solution was to use violations of laws (e.g. privacy, copyright, data protection). Next, a wide range of research literatures investigating information disorders guided the selection of risk factors to be included in a risk-based framework (e.g., target, actor, content, form, manner of distribution, consequences). Moving from the abstract to the concrete, the criteria and risk factors were examined with reference to current real-world examples of information disorders drawn from a variety of settings (e.g., academic research, advertising campaigns, cyberattacks, peer-to-peer corporate transactions, public weather warnings). One of the main challenges exposed through this process was evidencing impact (tangible losses, intangible losses). This, and other limitations that were revealed through real cases are discussed in the concluding section which stresses the need for an incident database of information disorders. Having a shared resource of real-world cases could better inform the development of a risk assessment methodology as well as advancing theory development and closing some evidence gaps through future empirical research.

Keywords: Risk assessment; Information disorder; Misinformation; Disinformation; Existential threats

Authors note: There was no financial support in preparation for this article. The author reports there are no competing interests to declare

********** Accepted Version To appear in Journal of Risk Research**********

Introduction

Information disorders (Wardle & Derakhshan, 2017) (e.g. misinformation, disinformation, fake news, conspiracy theories) are now referred to as existential threats (Ecker et al., 2025; Porter et al., 2025; Roozenbeek & Van der Linden, 2024; Van Raemdonck & Meyer, 2024). To appreciate why this is the case we need only consider the affects they have been associated with: scepticism and cynicism towards traditional news media (Markov & Min, 2022), eroding trust in government (Ahmed et al., 2025), societal polarization (Vasist et al., 2024), election interference (Echeverría et al., 2025; Ecker et al., 2025; Lin et al., 2025; Schroeder et al., 2026), poorer health (Paul & Yesmin, 2024; Van der Linden et al., 2025), undermining national security (Pierce et al., 2022), stock market disruptions (Arcuri et al., 2023), supply chain disruptions (Petratos & Faccia, 2023), economic instability (Asevameh et al., 2024), damage to critical infrastructure (Barker et al., 2025), and reputational damage of companies (Zhou et al., 2024).

Concern amongst public and private organizations is high because the job of developing control measures to tackle information disorders increasing in difficulty as more of their impacts are discovered. In fact, designing appropriate treatments for addressing information disorders, like all risk management, should start with risk assessments, to estimate the prevalence and severity of impact different information disorders pose (e.g., Aven & Thekdi, 2022; Pamment, 2022; Trammell, 2020; Varela da Costa et al., 2025).

Currently there is no general-purpose risk-based framework to inform risk management decisions to tackle information disorders, and so to address this the aim here is to propose some criteria and risk factors that could be included in a general framework. The criteria are used to prioritise the application of the risk framework according to the type of disorder (i.e. disinformation, hoaxes, fake news, conspiracy theories) and the domain (i.e. digital environments) to locate them in to. Finding convergence in research on the common properties associated with information disorders guided the set of risk factors to include in the framework (i.e. actor, content, form, manner of dissemination, target, consequences). The first two sections of this article were inspired by current research on the topic of information disorders that either use idealised scenarios, hypothesised examples, or some real-world examples to study the impact of information disorders. To explore the realistic challenges of evaluating the effects of information disorders, the third section uses current real-world examples to assess the viability of the criteria and risk factors proposed here. Several evidence gaps, and other limiting factors are discussed in the concluding section which also considers future directions and practical steps including the development of an information disorder incident database.

Designing a risk assessment framework for information disorders

Information disorders encompass any type of content that is distorted and that is distributed in ways that can cause a variety of harms. The distortions can take many forms (e.g. misleading content, false content, fabricated content, manipulated content, impostor content, false connections, opinions presented as facts, errors in facts, fabricated sources, satire or parody) (e.g. Kumar et al., 2025; Osman et al., 2022; Pérez-Escolar et al., 2023). To sort the distortions in ways that are meaningful one distinguishing feature is whether (e.g. disinformation) or not (e.g. misinformation) the distortion is premeditated in mind of producing a harmful outcome (e.g., Pérez-Escolar et al., 2023; Søe, 2021). Abundant are incidental falsehoods and inaccuracies, which some have argued are part and parcel of everyday communication (e.g., Andersen & Søe, 2020; Sperber et al., 2010). The deeper concern comes from willfully constructed and communicated falsehoods and inaccuracies, which aim to create discord, distrust, or material damage to individuals, groups, private and public institutions. Evaluating the impact of all types of information disorders may be difficult as well as impractical. The first step is to propose criteria to help narrow the focus on the type of information disorder to consider and where they occur to enable an evaluation of impact through a risk framework.

Criterion 1

Which types of information disorders to prioritize an evaluation of their impact: Disinformation is defined by some as “verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm” (Van Hoboken et al., 2019, p. 15). Along with this, there are an increasing range of information disorders that fall into the scope of being willfully designed to cause harm. Take for instance, fake news, which is generally referred to as the deliberate construction and dissemination of false, exaggerated

and biased news reporting to influence opinions and behavior (e.g. Gelfert, 2018). Propaganda combines truth claims to lend credibility with false claims that help to capture a mass audience (Walton, 1997). By design it contains ideologically slanted content to further a cause and to styme critical analysis (e.g., Henderson, 1943; Walton, 1997). Hoaxes are memetic because they also masquerade as true by combining truth claims and fabrications to confuse as well as interest audiences (Fleming & O'Carroll, 2010; Utami, 2018), and this is also the case for conspiracy theories (Douglas & Sutton, 2023). Conspiracy theories also subvert widely accepted assumptions of events, describe malevolent or forbidden acts, and ascribe nefarious ends to individuals or groups (e.g., Douglas & Sutton, 2023; Van Prooijen & Van Vugt, 2018).

As can be seen from the definitions of these types of information disorders (disinformation, fake news, propaganda, hoaxes, conspiracy theories) the common denominator is intentionality to cause discord, distrust, or material damage (Ó Fathaigh et al., 2021). For ease, these examples are collectively referred to as Willfully Deleterious Information Disorders (WDID).

While setting out which examples of information disorders to include, by extension there also needs to be consideration of what to exclude. WDID excludes examples of content that includes fabrications and falsehoods, but with the intent to amuse or to inspire critical thought, such as parody and satire (Adams et al., 2023; Kumar et al., 2025; Levak, 2020). Also excluded from WDID are examples of information disorders where the content is misleading but is generated and shared inadvertently, such as misinformation. Moreover, while negative outcomes from misinformation may occur, currently there is a need to improve the methods used to establish a reliable causal relationship between the two (Adams et al., 2023; Altay et al., 2023; Granados Samayoa, & Albarracín, 2025; Osman, 2025; Scheufele & Krause, 2019); though others would argue that there are sufficient grounds to evidence adverse outcomes from misinformation (e.g., Ecker et al., 2025; Roozenbeek & Van der Linden, 2024).

The basis on which to broadly distinguish examples that fall into and out of scope of WDID the main criterion is based on intent to cause harm, but other specifications are needed. For instance, it is clear from the various definitions of WDID that they can contain a mixture of truth claims as well as falsehoods and fabrications. Moreover, included in information disorders is mal-information which is true content (e.g., Adams et al., 2023), but can be distributed for malicious ends (e.g. exposing individual to harm by realising private information to the public). These added complexities pose problems for deciding *a priori* a process for identifying content that can be WDID. Also, risk analysis is not usually employed to arbitrate between truth and false claims (Lindaas & Bakken, 2025), because it is functionally designed to generate estimates for justified beliefs about different states of the world (Ylönen & Aven, 2023). This means that the criterion for distinguishing which types of information disorders fall in and out of scope of WDID needs independent adjudication.

One possible solution comes from legal studies (e.g., Chang, 2022; Li, 2025; Peukert & Windisch, 2025; van Hoboken & Fathaigh, 2021). Disinformation and other such examples of WDID are not illegal, at least not for now (Ó Fathaigh et al., 2022). However, there is specific content contained within examples of information disorders that does violates laws (e.g. privacy laws, copyright laws, defamation, public order) and breaches regulations and standards (e.g. advertising standards) (e.g., Chang, 2022; Li, 2025; Peukert & Windisch, 2025; van Hoboken & Fathaigh, 2021). Given this, as a practical solution, and avoiding painful debates about definitions, Criterion 1 incorporates this legal approach. The idea here is to apply independent formalised norms to assess content according to whether it is deliberately designed to cause harm given that the content itself violates laws, breaches regulations and standards referred to here.

Criterion 2

Criterion 2 considers the domain of distribution to also help focus an evaluation of the impact of information disorders based on where likely widest range of impacts can be caused. The digitisation of information means there are numerous forms that information disorders can take (image, video, audio, text, mixed media, 3D content), and that can exploit generative artificial intelligence (GenAI) tools (Kumar et al., 2025; Schroeder et al., 2026). Because digitized content is compatible with globally accessible media technologies (e.g., Mulcahy et al., 2024; Rathje & Van Bavel, 2025) strategies can be used to optimise WDID virality, i.e. accelerated speed and reach of target audiences. Furthermore, virtual environments arguably enable the exploitation of social mechanisms to amplify the impact of WDID by expanding audience reach more efficiently than offline mechanisms (e.g., Caramancion, 2020; Caramancion et al., 2022). Outside of media technologies, digitised forms of WDID can also permeate critical infrastructure (e.g., Barker et al., 2025; Jamalzadeh et al., 2024; Khameneh et al., 2025). Thus, exploiting digital environments has dual impacts, where harm can occur online, with profound spillover effects offline (e.g., Chadwick et al., 2025; Eriksson Krutrök, &

Lindgren, 2022; Kwon et al., 2024; McLoughlin & Brady, 2024; Schroeder et al., 2026). Thus, criterion 2 is used to prioritise the evaluation of the impact of WDID through digital environments, which excludes offline mechanisms. Not only because virality and amplification is arguably greater in digital environments than offline, but because there can be corresponding impacts online and offline which increases the range of impacts that could be observed.

Risk Factors

The process of building a risk-analysis framework to evaluate the impact of information disorders starts with two broad criteria for deciding on what and where to focus the evaluation. The next step is to look for candidate risk factors to include in the framework. To do this a wide range of literatures were considered (e.g. cognitive science, communication studies, computer studies, economics, legal studies, Information/Library Science, political sciences, sociology). This involved surveying several empirical studies and reviews (e.g., Asevameh et al., 2024; Bontcheva et al., 2024; Chechelashvil et al., 2023; Ferrara, 2024; Kumar et al., 2025; Li, 2025; Olivares-Delgado et al., 2022; Osman, 2025; Paul & Yesmin, 2024; Pérez-Escolar et al., 2023; Pierce et al., 2022; Plikynas et al., 2025; van Hoboken & Fathaigh, 2021; Vasist et al., 2024). The main purpose here is to cast a wide enough net to reveal properties of information disorders that different research communities converge on. The rationale for each of the five factors (i.e. target, actor, content, form, manner of dissemination, consequences) is presented in detail when discussing them in turn. The presentation of the five factors is organised in a specific way that corresponds to a basic structure of a risk assessment methodology (e.g., International Organization for Standardization 2019) so that the factors could grouped into those that are hazards (or threats) and vulnerabilities that contribute to causing harms (consequences).

Target

Risk Factor Target: A broad range of possible targets of WDID have been identified in recent literatures, either based on real examples, or by considering idealised scenarios (e.g., Arcuri et al., 2023; Asevameh et al., 2024; Chechelashvili et al., 2023; Harashima, 2023; Li, 2025; Olivares-Delgado et al., 2022; Paul & Yesmin, 2024; Pérez-Escolar et al., 2023; Rickards, 2024). Researchers have attempted to differentiate who the primary target of WDID from the wider range of secondary impacted entities. This is because there are cascading effects of WDID, which means several other targets end up being affected across time.

To illustrate, Arcuri et al. (2023) consider cases of fake news impacting stock prices in different stock exchanges (e.g. New York Stock Exchange (NYSE), National Association of Securities Dealers Automated Quotation (NASDAQ)). The fake news was distributed via social media and traditional news sources (e.g. business wires, newspapers) and targeted specific publicly listed firms (either national or multinational) that were named in the news stories. Those targeted, and other associated firms (i.e. secondary targets) were negatively impacted in terms of stock prices, the firm's earning from the stocks, and wavering investor confidence. Arcuri et al. (2023) also observed that there were incidental beneficiaries of the WDID because they were competitors of the targeted named firms.

Primary and secondary targets were examined by Kausa et al. (2024) this time when WDID was used to attack critical infrastructure (e.g. sewage plants, nuclear power stations, power grid). The WDID took the form of malicious insertion of false data into the control system. Several examples of this form of WDID involve destabilising the functions of systems that control sewage plants resulting in spillage of raw sewage into rivers, or power grids that resulted in a national blackout (e.g. Kausa et al., 2024). The primary target of the WDID was the specific control system, but the cascading effects observed as secondary targets were damage to the local ecology, or adverse impacts on the local population.

For the purposes of the risk assessment framework proposed here, the risk factor 'target' refers to the primary entity the WDID is designed to harm, disrupt, or cause material damage, while accepting that secondary targets will likely be impacted as well (e.g., Arcuri et al., 2023; Asevameh et al., 2024; Chechelashvili et al., 2023; Harashima, 2023; Li, 2025; Olivares-Delgado et al., 2022; Paul & Yesmin, 2024; Pérez-Escolar et al., 2023; Rickards, 2024). The primary target can include the following: high-profile individuals, a social group, a religious group, a political party, the public, a public institution, a corporation, national infrastructure, a nation state.

Actor

Risk Factor Actor: The first thing to establish is that the actor responsible for the WDID may be different from the actor that disseminates it. Candidate examples of responsible actors range from foreign state actors, hostile states, dark PR companies, cybercriminals, hacktivists, corporations, news outlets and well as

figureheads, celebrities and influencers (e.g., Caramancion et al., 2022; Chechelashvili et al., 2023; García-Ull et al., 2025; Hameleers, 2023; Olivares-Delgado et al., 2022; Vasist et al., 2023). This means that an important distinction needs to be made between a nefarious actor (e.g., dark PR company, hostile state, corporation) that designs WDID, from the witting as well as unwitting actors that distributes WDID. The witting actor distributing WDID is discussed under the risk factor 'Manner of dissemination'.

The actor generating the WDID content also needs to be distinguished from how artificial intelligence may be exploited to design or disseminate WDID content (e.g., Gabriel et al., 2024; Kumar et al., 2025). In the advent of GenAI and the tools available (e.g. Bard, Bing AI, ChatGPT, DALL-E, My AI) an artificial entity can be trained to generate fabricated, false, or misleading content (Bontcheva et al., 2024; Gabriel et al., 2024; Schroeder et al., 2026). Virtuous (or at least unintended) hallucinatory text-based and visual-based errors that are made (e.g. Liu et al., 2024; Ye et al., 2023) can result from the training data sets as well as the algorithms themselves, both of which independently and in combination lead to information disorders, but not necessarily WDID. In contrast, actors may generate training data that contains false information to support the generation of new false content, in a wilful manner (Bontcheva et al., 2024; Schroeder et al., 2026). Also, artificial agents (e.g. social bots) can be employed in malicious ways to communicate WDID (e.g., Hajli et al., 2022; Plikynas et al., 2025). Nonetheless while it is worth understanding the exploitation of AI technology, human agency is key. Therefore, for now, in the current proposed framework, the entity and their malicious motivations to exploit the technology, starts with a human entity which can be: a foreign state, hostile state, dark PR company, cybercriminals, hacktivists, corporations, news outlets, figureheads, celebrities and influencers.

Content

Risk Factor Content: WDID can be difficult to identify because the content contains a mixture of falsehoods, fabrications and truth claims (e.g., Ferrara, 2024; Olivares-Delgado et al., 2022; Paul & Yesmin, 2024; Pierce et al., 2022). Also, misclassifications can occur because the status of claims evolves over time, where claims purported to be accurate are later found to be false, but also highly contentious claims ascribed as false are latter accepted as true (e.g. Adams et al., 2023; Berkowitz, 2021). Recent classification systems designed to serve as independent authentications and verifications of information disorders employ digital book-keeping ledgers such as block-chain and digital forensics techniques (Ahmad et al., 2024; Arquam et al., 2021; Bennke, 2023; Polčák, & Kasl, 2021). As promising as this is, for now in the absence of a consensus on which authorisation and verification methods are best, an alternative independent norm is needed to help assign content into the category of WDID.

The approach proposed for Criterion 1, also extends specifically here. Adopting a legal framework for assessing whether content could violate privacy laws (e.g. Li, 2025) or copyright laws (e.g., Chang, 2022; Peukert & Windisch, 2025) serves as a independent norm. Privacy regulations, Privacy laws and Copyright laws in combination prohibit the ability for private or public actors to collect, use, store, transfer, or sell personal information (Li, 2025), as well as reproduce, distribute, adapt, display or lend content (e.g., original literary, dramatic, musical, and artistic works, films and sound recordings, broadcasts, databases, software, photographs, illustrations, and other images) (Peukert & Windisch, 2025). Therefore, as a risk factor, the content of WDID can be determined with recourse to evaluating the possibility that it violates privacy, copyright, or data protection laws, or other regulations and standards.

Form

Risk Factor Form: As discussed, there are numerous digital forms that WDID can take (e.g., Ahmed & Pathan, 2020; Caramancion et al., 2022; Kumar et al, 2025; Pooranian et al., 2021). By focusing on digital forms, this in turn helps to expand the evaluation of the impact of WDID to include cybercrimes (e.g. Caramancion et al., 2022; Shubham et al., 2025) and critical infrastructure (e.g. Ahmed & Pathan, 2020; Barker et al., 2025; Jamalzadeh et al., 2024; Khameneh et al., 2025). Table 1 summarises the different forms that WDID can take (e.g., Caramancion et al., 2022; Kumar et al, 2025; Pooranian et al., 2021). The dedicated section discussing impact "risk factor: consequences" goes into more detail, but Table 1 helps to contextualise how the forms of WDID are described by researchers as causing harm. It is worth recognising here that situations will arise that include combinations of digital forms of WDID, and which combinations are most likely to occur requires empirical work. For now, the starting point here is to detail which forms have currently been investigated, and from this future work can evidence new forms that exploit technological advancements along with typical combinations that are employed to cause harm.

Table 1. Techniques used for generation of wilful deleterious information disorders (WDID)

Technique Name	Description	Potential Impact
Badvertising	A kind of camouflaged click fraud attack on the advertising industry, automatically generates click-through on an advertisement when users visit the website in the form of a malicious mutation of spam and phishing.	The attack artificially and stealthily increases the number of clicks on ad banners hosted by the fraudster or unaware associates to generate more revenue for the attacker through advertising.
Clickbait headlines	Crafting sensational headlines to attract clicks, often with misleading or false information within the article.	Misleading readers and driving high traffic towards fabricated articles, or other false content.
Conspiracy Theories	Propagating elaborate and unverified descriptions of events as factual via news articles, videos, and other media channels.	Undermining trust in official narratives and promoting distrust through fabricated and/or false content.
Deepfakes	Using generative AI to fabricate video and/or audio to depict individuals saying or acting in ways that are entirely fabricated.	Creating convincing but false or fabricated multimedia content in order to dupe individuals, groups, or the public.
Fabricating Sources	Inventing or attributing false sources in fake news, fabricated articles, or other false content.	Lending false credibility to fake news, fabricated articles, or other false content.
False data injection attack (FDIA)	The introduction of manipulated or fabricated data into a system, specifically targeting the data used for state estimation or control decisions in critical infrastructure.	Operational problems, instability, or even physical damage of critical infrastructure (e.g., power grids)
False Expertise	Falsely claiming expertise in a subject to present false information as factual or false attribution of expertise to claims being disseminated.	Lending credibility to fake news, fabricated articles, and publicly disseminated opinions.
Impersonation	Falsely presenting content as attributable to a reputable individual or institution.	Duping individuals, group, or the public into thinking that the false/fabricated content of communication is from a trusted source.
Malware	Viruses, worms, trojans; viruses are executable programs that insert codes into legitimate programs. Worms are self-replicating programs that spread in systems to drain their resources. Trojans are malicious programs disguised as legitimate software aimed at damaging a system.	Stealing personal information (i.e., financial or health); falsify or modify personal data; lock access to or release sensitive information to the public.
Malvertisement	Designing a platform for distributing malware by injecting malicious code (e.g. worms, viruses, trojans, botnets) into legitimate ad networks.	Malicious ads take advantage of browser vulnerabilities to infect the victim's system, persuade users to download and to install malicious software
Manipulated Media	Distort images, videos, audio, maps, 3D content, Simulated environments in order mimic genuine media content.	Duping individuals, groups, public, into accepting false or fabricated content as real.
Misquoting & Out-of-Context Quotes	Altering or taking statements out of context to change their meaning. Using statements made in one context in a different context to change their meaning.	Mispresenting the views of public figures in order to lend credibility to false or fabricated content.
Phishing	Fake emails, fake SMS or instant messages, and fake websites that may look authentic	Disruption of system operations; alter, damage, steal, or disrupt data
Ransomware	A Trojan or a worm is deployed via phishing or visiting a compromised website, where malicious software installs on a system or computer, causing that system or information to be encrypted. Upon encryption, a ransom message is displayed stating the deadline for monetary payment (often in bitcoin).	Expose sensitive, personal, or embarrassing information unless ransom is paid

Manner of Distribution

Risk Factor Manner of Distribution – communication paths: Several reviews have catalogued the way WDID is disseminated online (e.g. Broda & Strömbäck, 2024; Burcă-Voicu et al., 2025; Oksanen et al., 2024; Plikynas et al., 2025; Ruiz, 2025; Wells t al., 2024). The list is vast, and includes: social media platforms, mainstream media outlets, alternative media outlets, media outlets that masquerade as legitimate news outlets, websites, blogs, and vlogs, digital marketing, online gaming, e-mail marketing, digital advertising, social bots, microtargeting, consumer reviews. The virality and amplification of WDID via digital environments means the dissemination paths of WDID can also change over time. For instance, a fake news story could initially appear as an article on a personal website, which is then posted on a social media platform, that could end up being viral through amplification by exploiting social bots (Doshi et al., 2023). Contrast this with examples of a more localised mechanism of dissemination for an entirely different form of WDID. For instance, an FDIA exploits a wireless communications network device that uses 5G to disseminate corrupted and/or false data in a control system of a critical infrastructure such as a power grid (e.g. Ahmed & Pathan, 2020; Liu et al., 2024). Where the FDIA is initially injected is separate to where else it starts to have cascading effects along the way it is carried across the control system (e.g. Ahmed & Pathan, 2020). Therefore, as a risk factor the manner of dissemination needs to be sufficiently broad to accommodate all these examples, and at the same time distinguish between primary, secondary and even tertiary methods of distribution over time.

A way to capture the different methods of dissemination is to look to communication theories on

networks and communication paths (e.g. Cappuccio et al., 2021; Fawkes & Gregory, 2001). We can think of the distribution of WDID based on whether the initial node in the communication path is human in origin, artificial agent, algorithmic, or a combination (e.g., Metzler & Garcia, 2024). From this, the communication path can take one of several structures which reflects different strategic methods of dissemination. This can be one-to-one (e.g. such as FDIA in which a packet of corrupted data is inserted into a system), one-to-a-specific group (e.g. a single deceptive email including malicious links that is distributed to staff within an organisation, or a false advertisement through microtargeting that reaches multiple consumers), one-to-a-organisation (e.g. a deepfake of a high ranking official sent to an organisation), one-to-a-network (e.g., a high profile influencer makes false claims to their followers on social media platforms), one-to-a-community (e.g. AI generated content contained in emails targeting a special interest group), and network-to-network (e.g., artificial agents such as social bots distributing AI generated posts on a social media platform). This is not exhaustive, but it sets out an approach for capturing the dissemination path and can, in future iterations, be used to capture the dynamics of dissemination of WDID over time.

Consequences

Risk Factor Consequences: One obvious difficulty in identifying the consequences of WDID is that the intended outcomes may differ from the outcomes that have been observed, not least because the WDID themselves may not be detected immediately. For instance, the difficulty in assigning the consequence of hoaxes is because only years later the originator announces the hoax or only much later, as is the case with examples of Wikipedia, has it been possible for hoaxes to be debunked (e.g. Borkakoty & Espinosa-Anke, 2024; Young, 2017). Moreover, given the dynamic nature of the distribution paths of WDID, identifying associated consequences needs to factor in appropriate time scales for their detection. For instance, when WDID were designed to cause discord in religiously fractious communities, after they were distributed the observable impacts occurred some days after or several months after (e.g. Paul & Yesmin, 2024). Others have documented both short term (e.g., days) and long-term (e.g. years) macroeconomic (e.g., Harashima, 2023; Rickards, 2024) and microeconomic damage (e.g. Gavurova et al., 2024) that WDID can cause, such as increased speculation in global financial markets.

For now, it is hard to find a principled basis on which to define an appropriate timeline to assess impact of WDID, therefore in the absence of this, the focus here is to simply consider which consequences of WDID are commonly reported on in the literature. One issue is that several studies examine hypothetical as well as actual examples of consequences of WDID (e.g. Ahmed & Pathan, 2020; Asevameh et al., 2024; Caramancion et al., 2022; Chechelashvili et al., 2023; Ferrara, 2024; Kumar et al, 2025; Li, 2025; Olivares-Delgado et al., 2022; Paul & Yesmin, 2024; Pérez-Escolar et al., 2023; Pooranian et al., 2021; Vasist et al., 2025). Some relate WDID to economic and financial losses, data theft, structural damage, public disorder, disruptions to services, physical injury, and loss of life. More complex consequences of WDID include erosion of public trust in democratic functions, reputation damage of individuals, groups, organisations and public institutions, distortion of public opinion, polarization. To simplify the process of capturing these various impacts, they are grouped into two: Tangible losses (fatalities and injuries, economic losses, financial losses, structural damage) and Intangible losses (cultural losses, social losses, political losses, reputational losses).

General-purpose risk-analysis framework of WDID: Accepting the many outstanding issues and limitations that have been discussed thus far, this first section was guided by current research to propose several risk factors that could be included in a risk-analysis framework of WDID, which are summarised in Table 2.

As mentioned earlier, the order of presentation of the risk factors was in mind of how they might be organised along the lines of typical risk assessment methodologies. Figure 1 shows how the risk factors can be grouped into those that would constitute threats (i.e. intent, capabilities), vulnerabilities, and overall impact. That is, the path starts from intent to cause harm, followed by the means by which harm is caused, the weakness that enable the harm to be caused, and the types of harms that are expected to occur. For instance, the target and the actor are factors that could be construed as indicators of intent since they imply the underlying motives of WDID. The content and form WDID takes might suggest the type of capabilities needed for their construction, such as how much planning and skills are needed. The manner in which WDID is distributed reflects the vulnerabilities in digital environments that are exploited. Thus, by ordering the risk factors along the lines of intent, capabilities and vulnerabilities we can nominally trace a causal path because causes of harms and their effects (Fenton & Neil, 2013; Neil et al., 2021).

Table 2. Summary of Risk Factors included in a Risk assessment framework of WDID

Intent		Capabilities		Vulnerabilities	Impact
Target	**Actor**	**Content**	**Form**	**Manner of Distribution**	**Consequences**
high-profile individual, a social group, a religious group, a political party, the public, a public institution, a corporation, national infrastructure, a nation state	Individual, high-profile individual, foreign state actors, hostile states, dark PR companies, cybercriminals, hacktivists, firms, corporations, news outlets	Premeditated design and distribution for the purpose of causing harm. Number of, and type of violations laws (e.g. Privacy Laws, Copyright Laws)	Badvertising, Clickbait headlines, Conspiracy Theories, Deepfakes, Fabricating Sources, False data injection attack (FDIA), False Expertise, Impersonation, Malvertisement, Manipulated Media, Misquoting and Out-of-Context Quotes, Phishing, Ransomware.	One-to-one, one-to-specific-group, one-to-an-organisation, one-to-network, one-to-a-community, network-to-network	Tangible losses (fatalities and injuries, economic losses, financial losses, structural damage). Intangible losses (cultural losses, social losses, political losses, reputational losses)

The proposed risk-analysis framework is based on common properties of information disorders that have been individually investigated in studies on WDID. Many gaps in the literature remain as discussed throughout this section. For instance, we currently do not know which combinations of forms of WDID are associated with particular types of impacts. As yet, it is hard to establish a temporal window for assessing the way WDID to track how their dissemination changes over time, and how the impacts may change over time. Also, the different forms of WDID may need to be weighted differently to reflect their differential impacts, which as yet is hard to determine. Finally, it may be the case that a general-purpose risk-analysis framework is too crude to adequately accommodate the way WDID are employed to cause harm in different contexts (e.g. to disrupt critical infrastructure, provoke social unrest, cause reputational damage to public institutions).

Accepting these many limitations, what has been presented here is simply a skeletal structure of how risk factors could be organised, along with the criteria for where, in principle, a risk-analysis framework could be applied to assess the impact from WDID. Clearly a lot more needs to be done for the translation of the proposed framework into a formal risk assessment methodology of WDID. Thus far what has been discussed in this section considers a risk-analysis framework in the abstract, and so the next section explores the viability of the criteria and the risk factors using real world cases.

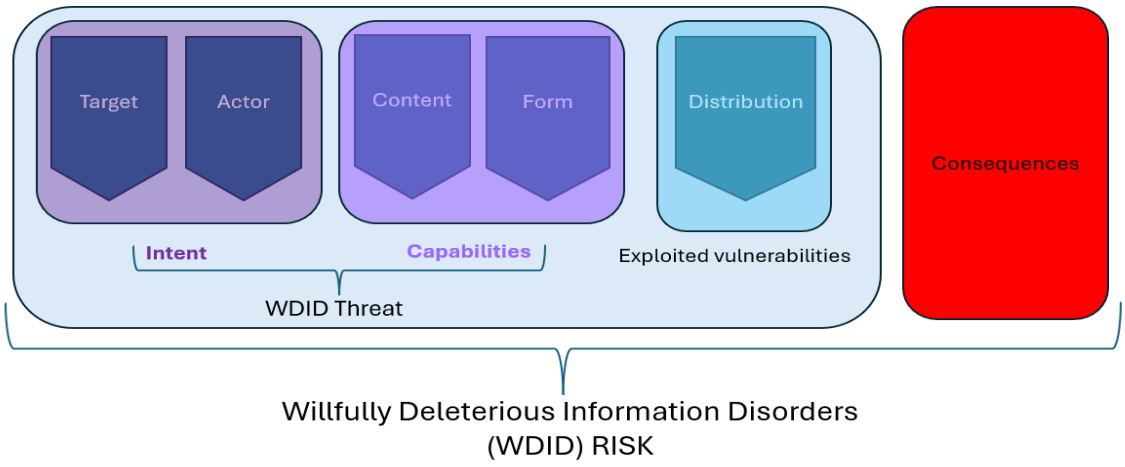

Figure 1. Risk-analysis framework for WDID including five risk factors (target, actor, content, form, distribution method)

Table 3. Cases of wilfully deleterious information disorders (WDID) between January 2024 to February 2026

	Context and link to news articles reporting the story	Details	Target	Actor	Content and Form	Distribution method	Consequences
1	Social media platform (Hal, 2025; Yousif & Jamali, 2025)	17/07/2025. Facebook [now META] claimed on its website that it had invested billions of dollars into protecting user privacy since 2019. On 17/07/2025 Facebook settled out of court, facing $8 billion in damages for their prior misleading and false privacy claims. In addition, the Federal Trade Commission (FTC) fined Facebook $5 billion in 2019 for failing to protect user data, as part of Facebook's 2012 agreement.	Facebook users	Facebook	False claims of data protection and data privacy. Breaching Privacy Laws.	Organisation-to-network. The organisation provides details of privacy and data protection on their website and on the social media app.	~13 billion USD in fines.
2	Weather warning (Mascelliono, 2024; NCSC, 2024)	14/11/2024. MeteoSwiss is the Federal Office for Meteorology and Climatology. Fraudsters sent letters to Swiss citizens disguised as MeteoSwiss. Recipients that were targeted were instructed to download a new "severe weather warning app" via a QR code contained in the letter. In actual fact it was a phishing scam that installed malware on smart phones in an attempt to steal sensitive data.	Swiss citizens	Entity unknown	Impersonation of an institution, Phishing attack. Fabrication of a weather warning app. Breach of copyright laws.	One/Group-to-targeted group. Physical letters distributed to targeted individuals that contains a QR code to download as an app on their smart phone.	The number of victims of the scam has not been documented.
3a	Cyberattack warning (Greig, 2025; MTN Group, 2025)	25/04/2025. South African telecommunications company MTN Group is the largest in Africa. MTN warned their customers that an "unknown third-party" had claimed to have unauthorized access to personal information of some customers. While there was a cyberattack, MTN had issued a statement that there was no indication that customer accounts or mobile wallets had been directly affected, though media reports indicate that some customers were notified of possible concerns with their data.	Customers and Investors in African markets	"unknown third-party" cyberattack group	False claims of having accessed critical data. Potential breach of privacy laws.	Unknown-third-party-to-public. Unknown third-party communicates publicly about have committed a cyberattack.	The number of customers affected is unspecified.
3b	Cyberattack warning (Chalk, 2024; NME, 2024)	27/02/2024. Mogilevich hacker group claimed on the dark web that they had launched an attack on Epic Games's servers and had accessed internal emails, passwords, payment information and source code that it was offering for sale. Epic Games reported that the claims by Mogilevich were in fact false, which was later confirmed by members of the hacker group.	Epic games, an American video game, software developer and publisher.	Hacker group	False claim of a cyberattack breaching security and stealing data. Potential breach of privacy laws	Hacker Group-to-network. Posting on the dark web.	No evidence of theft of data, or compromised systems.
3c	Cyberattack (Butler, 2025; Musafer & McMahon, 2025)	20/04/2025. Marks & Spencer (M&S) experienced a breach of their IT services, which was suspected to have occurred in February 2025, but was publicly announced in April. The hackers impersonated an M&S employee convincing support staff to reset the employee's password and hand over access to internal systems. This help-desk manipulation gave the intruders valid credentials to M&S's corporate network, bypassing perimeter defences without resorting to malware or exploits. The installation of their malicious software code to effectively scramble dozens of virtual machines and bringing critical applications to a standstill.	Marks and Spencer British Multinational retailer	Hacker group	Impersonation of staff. Phishing attack. Insertion of malware. Breach of copyright laws.	Group-to-network. The hacker group use a single point of entry via the IT helpdesk to then exploit internal IT company systems.	~650 million USD loss of market value
4a	Academic Profile (Andrei, 2024; Wilcox, 2024)	22/07/2024. Using Generative AI to fake scientific articles. To illustrate the success of this Dr Reece Richardson an academic	Entity unknown	Academic	Fabricated scientific articles using generative AI tools to produce	One-to-network.	Unspecified.

		at Northwestern University USA generated false academic articles uploaded on Research Gate a social media platform for academics. The attributed author of the articles is Larry Richardson, who in fact Reece's grandmother's cat.			content. Impersonating an academic by fabricating an academic profile Breach of copyright laws.	Designing an academic profile online with content that was automated generated.	
4b	Academic Research (Ibrahim et al., 2025)	14/02/2025. Ibrahim, Liu, Zaki and Rahwan (2025) published an article exposing how online services provide to academics for a fee as "H-index & Citations Booster". These services are used to falsely inflate academics credentials and prestige.	Academia	Individuals and companies offering and/or brokering the services	False promotion of academic credentials, False generation of citations. Potential breach of privacy laws and copyright laws.	One-to-network. e-mail, webpages,	Unspecified.
5	Music Band (Kleinman, 2025; Maimann, 2025)	05/07/2025. A Canadian resident using the pseudonym Andrew Frelon posing as a spokesman for the band Velvet Sundown revealed to journalists that he used generative AI platform Suno to create the songs in an elaborate Art hoax. The band, the music and the promotion material that had been placed on social media and music distributions sites were falsely generated using AI tools.	Audiences using Spotify and other music hosting platforms.	Entity unknown	Fabricated music band, fabricated music, fabricated promotion material. Breach of copyright laws. Potential breach of Privacy laws.	One/group-to-public. Generative AI platforms for generating music content, image and voice manipulation software.	Financial impact on the music industry hard to determine, sales of music content, and duping over 850,000 Spotify listeners.
6	Corporate Transaction (Leng & Ho-him, 2024; Milmo, 2024)	29/01/2024. Hong Kong police received a report of a staffer in HK working for Arup (multinational company) that they had made 15 transfers of Arup company funds to five different Hong Kong-based bank accounts. The employee had been contacted first by email. After receiving a false email from a company's chief financial officer, a video conference call was arranged. The employee then entered a faked video conference call with several senior officers of the company requesting money transfers to designated bank accounts totalling 200million Hong Kong Dollars (~25 million USD).	multinational company	Entity unknown	Impersonating senior officials of a company. Deepfake and phishing attack. Fabricated email, and fabricated voice and images of senior company officials. Breach of copyright laws. Potential breach of Privacy laws.	One/group-to-one/company. Email and voice conferencing combined with image and voice manipulation software.	~25 million USD
7	Stock Market Trading (Krishna, 2024; Upadhyay, 2024)	01/04/2024. National Stock Exchange (NSE) of India LTD cautioned investors of a deepfake videos that had been distributed. The deepfake was of the CEO Ashishkumar Chauhan providing false stock recommendations.	Private investors, public	Entity unknown	Fabricated video and audio impersonating the NSE CEO. Deepfake. Breach of copyright laws.	One-to-network. Video and audio distributed via social media	Unspecified.
8	Dietary supplements (Mendelson, 2025; Wong, 2025)	07/03/2025. The Việt Nam Food Safety Authority under the Ministry of Health cautioned consumers about a new dietary supplement called 'Supergreens Gummies'. The concerns surrounding the false claims of the benefits of the supplement. The authority took action against celebrities and social media influencers who have been promoting this product on social media platforms like Facebook and TikTok.	Social media followers, public	Celebrities and social media influencers	False and exaggerated claims of effectiveness of food supplements. Breach of advertising standards.	Multiple individuals-to-network. False claims communicated on social media platforms.	30,000 customers compensated for their purchases totally 65K USD, and 9K USD in fines issued to the company of the products.
9	Foreign policy communications (Cook, 2025; Yousif, 2025)	08/07/2025. An unknown actor" was alleged to distributed manipulated media to artificially generate the voice of US Secretary of State Marco Rubio to contact three foreign ministers via the Signal messaging app, and leaving voice mails on the app.	US Secretary of State Marco Rubio	Entity unknown	Deepfake. Impersonation of a state official, false policy information. Breach of copyright laws. Potential breach of Privacy laws.	One-to-a-specific-group Image and voice manipulation software. Voicemails on social media app and text messages	Unspecified.
10	Terrorist attack (Barrett-Peters, 2025; Butler, 2025)	10/03/2025. Police were tipped off that there was a suspicious caravan in north-western Sydney caravan. It was found containing enough explosives to produce a 40m-wide blast and contained a note with a hate-group messages. An investigate found that this was a false terrorist plot orchestrated by a criminal gang to	Australian Federal police	Organised criminal gang	False construction of a terrorist attack. Potential breach of Privacy laws.	Group-to-targeted group(?). Planting physical documents alongside explosive materials to be discovered by law enforcement.	Unspecified.

		negotiate reduced sentences for those involved in plot. The potential terror threat was known to the public 10 days after, when the news was leaked to a Sydney newspaper.					
11	Weather Crisis (Chan & Sobhan, 2025; Sobhan & Chan, 2025)	29/10/2025. Hurricane Melissa was a powerful tropical cyclone this hit Jamaica as a hurricane in October reaching peak devastation between the 25th and 27th of October 2025. Around this time digital altered as well as entirely AI generated videos of the devastation where being circulated on social media. The videos depicted sharks swimming in a hotel swimming pool, as well as views of the eye of the hurricane from the porthole of a plane caught inside the hurricane.	Social media followers, public	Entity unknown	Deepfakes employing AI tools like OpenAI's Sora to distort or entirely fabricate the depiction of the weather event. Potential breach of privacy laws and copyright laws.	Multiple individuals-to-network. Video and audio distributed via social media	Unspecified.
12	Government Minister (Cimili, 2026).	12/02/2026. In September 2025 Albania introduced the world's first artificial intelligence minister in September 2025. This was when Prime Minister Edi Rama unveiled the AI-generated "minister" named Diella during the formation of his new government. On the 12th of February 2026 Anila Bisha a well-known Albanian film and theatre actress triggered a legal dispute over the use of her likeness for Albania's AI-generated virtual minister.	Public	Albanian Government	Deepfake. Impersonation of a public figure. Potential breach of privacy laws and copyright laws.	One-to-network. Video and audio distributed multiple official channels including social media, and state websites and applications.	Unspecified.
13	City Council communications (White, 2025).	March to October 2024. During 2024 city councillors in Canada were targeted with emails along with reports, articles and presentations containing AI generated content using a chatbot ""Canadian Civic Advisor," to distribute fabricated content around climate protection policies designed to influence decisions regarding net-zero	Public Officials	KICLEI ('Kicking International Council out of Local Environmental Initiatives') political activism movement	AI generated content using a chatbot. Potential breach of privacy laws and copyright laws.	one-to-a-specific-group KICLEI had targeted Canadian public officials through emails.	Voting behavior of city councillors
14	Multinational accounting and consultancy firm (Dhanji, 2025; Kissin, 2025)	6/10/2025 Deloitte issued a refund to the Australian Government for a report they had prepared which cost $440,000. Deloitte had originally failed to disclose the use of AI in preparation of the report, and the report itself had included several errors and fabricated references to scientific articles, and false quotations. In fact later in the year (25/11/2025) the Deliotte had alleged to have done the same for a report commissioned by the Canadian Government.	Government	Deloitte	Ai generated fabricated content. Potential breach of privacy laws and copyright laws.	One-to-One Deloitte provided the report containing fabricated content to the Australian Government.	The precise refunded amount has not been released to the public.
15	European Logistics Firm (Mitrovic, 2025).	2024. A European logistics firm (details undisclosed) that handles transport, was revealed to have been attacked through a data-poisoning method. It altered the firm's AI-training data sets in such a way as to reduce the efficiency of scheduling vehicles, fuel costs and delivery times. The net effect of this malicious attack severely impacted the operations of the firm.	Private sector organisation	Entity unknown	False data injection attack (FDIA) into the firm's AI-based route-optimisation model. Potential breach of privacy laws and copyright laws.	Multiple individuals(?)-to-Network The AI system was breached, and false data and manipulated data was introduced in the AI-model.	The firm experienced an estimated $4.7 million USD in losses from the attack.

Test Cases

The aim of this section is to use cases of potential WDID (see Table 3) to explore the realistic challenges when applying concepts relevant to a general-purpose risk-based framework. Some simple conditions were applied to select the real-world cases. First, cases were selected that were in the public domain, and so where possible traditional news sites were used to source them (e.g., ABC news, BBC, Boston Globe, CBC news, CBS, CNN, Financial Times, The Guardian, NME, PBS, Reuters, Yahoo news). The aim also was to explore cases that varied in context (e.g., political, social, entertainment, education) to explore the feasibility of having a general-

purpose risk-analysis framework. To ensure their relevance, and topicality, the search for cases was restricted to articles published between January 2024 to February 2026. To be inclusive, where possible cases were drawn from countries outside (e.g., Albania, Caribbean, Hong Kong, India, South Africa, Vietnam) as well as within countries often studied in research on information disorders (e.g., US, UK).

Each case in Table 3 is presented according to the context and citations, the date the cases was made public, along with a brief description. For the most part, the publicly available details of the cases were used to infer the target, actor, along with some basic descriptions of the content and form, manner distribution and possible tangible consequences. Table 3 includes cases where important details were missing, this is because the details are unknown or have not been made known to the public.

Application of Criterion 1 and 2: Cases 1, 3c, 4, 5, 6, 7, 8, 9, 11, 13 and 15 fulfil criterion 1 and 2, because they were distributed on digital environments, and the information disorders were deliberate in nature, appear to be designed to cause harm, and violated laws or various standards and regulations. However, other cases exposed several important issues concerning how well the criteria accommodated the particulars of the cases, and where there may be reason to revise them. For instance, case 2 involved the distribution of an information disorder via physical letters that impersonated a legitimate institution but contained a QR code that would distribute malware digitally. Given the malicious intent behind it and the potential laws it violated (e.g. copyright), this would meet criterion 1, but did not meet criterion 2 because the cases involved the distribution of WDID via physical letters. The features of this case are nonetheless interesting because while the WDID was distributed offline, the harm would have been caused via digital means based on the QR code the letters contained. Consider case 12 which refers to a situation of a fabricated public official, thus meeting criterion 2. However, in this case the public are aware that the official is AI generated so this would not necessarily meet criterion 1, though the issue with this case is whether the fabricated image breaches copyright laws. Also, consider case 14, where there is some ambiguity with respect to fulfilling criterion 1 but does fulfil criterion 2. With regards to the latter, the report itself was distributed as a digital report to government officials, likely over email. The report that was prepared by the multinational firm included AI-generated content which was found to be entirely fabricated (e.g. quotes, academic literature), though this does not in itself reflect a deliberate attempt to deceive. However, failing to explicitly disclose that some of the content was AI-generated, of which some was likely to be fictitious given poor efforts to quality assure the document, has been treated as failing to fulfil their legally contracted obligations. For this reason, this case is included for analysis.

The two cases (case 4a and 4b) concerning academia are also a useful stress test of the two criteria. Case 4a describes a situation where the academic had fabricated an academic profile, and because it was disseminated in a digital environment this would meet criterion 2. By impersonating an academic it could be seen as a breach of some laws or regulations thus meeting criterion 1. However, the intent behind the design of the fake prolife was to expose concerning issues with academic fraud. Similarly, case 4b also meets criterion 2. Case 4b described how academics can easily generate and publish fabricated articles as well as inflate metrics that index their productivity and impact by via citation cartels (Ibrahim et al., 2025).

The online distribution method of the fabricated profiles and fabricated academic means that case 4b meets criterion 2. As with case 4a, some aspects of the case 4b would meet criterion 1 because the profiles and articles were fabricated and likely violated copyright laws because both were associated with academic institutions to lend credibility. However, criterion 1 also requires that the WDID is designed with the intent to cause harm. Clearly the activities that were documented by the academics (Ibrahim et al., 2025) were for the purpose of illustrating how actual academic fraud can occur but did so by using fraudulent methods to expose the mechanisms available to academics.

Other examples of complex edge cases were 3a, 3b and 10. Cases 3a and 3b were false claims of cyberattacks and case 10 was a false terrorist attack. In all three cases then, the content contained false claims of an attack, but as false flags they still appear to be designed to cause harm or material damage, and so the content itself would meet criterion 1. Cases 3a and 3b meet criterion 2, while case 10 fails this criterion because the false claims were distributed offline, though over time the details of the case were later disseminated in mainstream media online and offline and were shown to cause considerable harm to the religious groups targeted.

The issues raised here are discussed in more detail in the concluding section, and for the remained of this section cases 4a, 4b, 10 and 12 were not discussed in any depth given that they failed to meet at least one of the criteria. Therefore, the remainder of this section focuses on cases 1, 3a, 3b, 3c, 4, 5, 6, 7, 8, 9, 11, 13, 14

and 15 to examine other challenges they present with respect to the five risk factors.

Targets and Actors

Real world Cases: Targets and actors: From the details presented in Table 3, it was possible for most cases to identify the target of WDID. Consistent with empirical work discussed earlier, the target of WDID can vary considerably ranging from a high-profile individual (case 9), social media users (case 1, 5, 8, 11), public officials (case 13, 14), multinational companies (case 3b, 3c, 6, 15), and public and private investors (case 3a, 7). While for most cases the primary target of the attack could be identified, there are likely to be secondary targets, though identifying them is difficult because they are still being investigated. What was more difficult was identifying who the actor was behind the WDID as can be seen in Table 3, where for the majority of cases the actor could not be identified. Again, one reason for this is that ongoing investigations mean that either the details are unknown or are yet to be made known to the public.

Along with this, complex cases such as 5, indicate ambiguity in who is responsible where the hoax was declared by an individual actor, but this was later denounced. Also, for some of the cases involving groups, e.g. hackers (case 3a, 3b) or entities that had may be individuals or hacker groups, they were often publicly referred to as "unknown entities". What these cases reveal is that the entity responsible for WDID can be elaborate, especially when it comes to designing cyberattacks (case 3a, 3b, 15), deepfakes (case 6, 9, 11), or impersonating staff to gain access to IT systems (case 3c).

Content, Form and Distribution

Real world Cases: Content, form, and distribution: In passing criterion 1, the cases considered here (cases 1, 3a, 3b, 3c, 4, 5, 6, 7, 8, 9, 11, 13, 14 and 15) indicate examples of various laws that had been violated, which was also used as a condition for determining the content of WDID as a risk factor. The aim here is not to present a detailed legal analysis, but merely to show that in principle examples of WDID can be assessed based on content that flout laws. While this could be done, it is important to highlight here that some judgment was needed to identify which properties of the WDID need to be considered against possible laws violated. For instance, in several cases (3c, 6, 7, 9, 11, 14) the content of the WDID involved the impersonation of an individual or an institution, which would imply violation of copyright laws. Case 1 while an example of privacy violations may be considered an edge case. The information published on the platform was genuine regarding privacy protections, but here it was the actions of the social media platform that were in breach of the privacy laws they claimed they were compliant with. As well difficult cases such as this, depending on how expansive the range of laws that could be violated, some cases suggest wire fraud (case 6), identity theft (case 3c, 6, 7, 9), breach of advertising standards (cases 5 and 8), data protection (case 15), and violating database rights and artistic rights (case 5 and 15). Taken together, outside of privacy and copyright, the cases reveal several other violations of laws and standards that could be included to assess the content of WDID.

There was enough variation in the examples selected to explore a range of forms that WDID could take, which indicate varying levels of capabilities. To illustrate, case 5 took multiple forms that involved the fabrication of music content, band images, and even quotes from the fictitious band members. Case 9 involved an audio deepfake as well as a false social media account, and false messages sent from the account. In fact, we see several cases (5, 6, 7, 9, 11) taking the form of deepfakes along with other forms such as phishing emails and messages (e.g. Case 6, 9 and 13). Case 8 is a more simplistic and straightforward WDID in the form of false and exaggerated claims concerning food supplements. Also, while sophisticated in form, the success of 3c cyberattack was contingent on falsely impersonating a staff member during a call with the IT help desk of the company. Taken together these cases highlight that more often than multiple forms of WDID are employed in the pursuit of causing harm.

Regarding the distribution mechanism, all the cases considered in depth here (cases 1, 3a, 3b, 3c, 4, 5, 6, 7, 8, 9, 11, 13, 14, 15) exploited digital environments. In some cases where social media platforms where the distribution method, networks were exploited, though the point of origin as a single entity, or a group was hard to identify (cases 3a, 3b, 5, 7). In some cases the path was one-to-one (case 6), but also one-to-a-specific-group (case 9, 13, 14), organisation-to-network (case 1), a group-to-one (3c), and multiple individuals-to-network (case 8, 11, 15). Overall, the cases here help to illustrate the range of networks and communication paths that are employed.

Consequences

Real world Cases Consequences*:* One important issue that had been discussed when considering the impact of WDID was the need to establish an appropriate temporal window. The current literature made it hard to establish a precise time frame. By examining real cases, the temporal issue clearly comes into sharp focus. Throughout the discussion of the cases, extracting important details to assess the viability of the risk factors was made difficult because they are still being investigated or the details are not made known to the public (e.g., case 1, 3c, 9). Because of this, using only the details currently reported means that not all consequences of the WDID are going to be accounted for. In addition, even focusing on reported tangible consequences for which some of the details are in the public domain (case 1, 3c, 6), in other cases the information has yet to be publicly disclosed (case 3a, 3b, 4a, 5, 7, 9, 14). For instance, loss of reputation could be judged according to loss in market value (case 3c, 14), directly incurred loses (case 15), or financial losses based on fraudulent activities (case 6). Case 1 indicates the financial penalties that the social media platform incurred, and for case 14 the firm was required to cover a refund (amount undisclosed). This raises the issue of whether the consequences to a company associated with WDID should be differentiated from the consequences (tangible or intangible) that are experienced by the victims of WDID.

When it comes to intangible losses, such loss of trust (case 11) and confidence in institutions (case 13, 14), in the absence of any metrics to gauge public perceptions and experience for each of the individual cases discussed here. These intangible consequences, at least based on the details presented in reports of the cases, can only be assumed rather than evidenced. For instance, with regards to case 13 where public officials were targeted with the aim of influencing their voting behavior, it would be difficult to ascertain the scale of influence. Not all public officials that were targeted would necessarily reveal whether key voting on climate protection policies were swayed positively or negatively by the fabricated content they received.

General Discussion

The fact that information disorders are treated as existential threats, and the need to find appropriate methods for tackling them arguably necessitates a risk-analytic approach to establish a principled way of estimating their impact. Without a framework to assess their impact, it is difficult to design appropriate treatments, not least because how they occur and where they occur varies substantially. Thus, to make some practical inroads, the aim here was to start the process off for designing a risk-analysis framework. This was done with full appreciation of the ongoing debates as to the precision of the definitions of different types of information disorders, concerns over the empirical methods to investigate the relationship between them and their effects, and what are effective methods of mitigation (e.g., Adams et al., 2023; Altay et al., 2023; Osman, 2025; Scheufele & Krause, 2019; Yee, 2023).

Thus, the starting point of this article was to focus on broadly accepted claims and assumptions that were taken at face value. Along with this the net was case wide to include research from a range of disciplines studying information disorders (e.g., Asevameh et al., 2024; Bontcheva et al., 2024; Chechelashvil et al., 2023; Ferrara, 2024; Kumar et al., 2025; Li, 2025; Olivares-Delgado et al., 2022; Osman, 2025; Paul & Yesmin, 2024; Pérez-Escolar et al., 2023; Pierce et al., 2022; Plikynas et al., 2025; van Hoboken & Fathaigh, 2021; Vasist et al., 2024). The objective here was to consider what a general-purpose risk-analysis framework would look like if it attempted to accommodate the different disciplines that study information disorders. In fact, because of this, a novel approach was adopted from legal studies as a way to solve definitional issues. By doing this, the content of WDID were classified based on the laws that they would be potentially violating (e.g., Chang, 2022; Li, 2025; Peukert & Windisch, 2025; van Hoboken & Fathaigh, 2021).

The rationale for looking at real-world examples was in view of proposing a framework that was practical, but also as a useful stress test to expose important weaknesses in the proposed risk-framework and the criteria. With regards to criterion 1 the case studies suggested some constructive insights for applying a legal framework to help classify the content of WDID (e.g., Chang, 2022; Li, 2025; Peukert & Windisch, 2025; van Hoboken & Fathaigh, 2021). The cases revealed that beyond privacy and copyright laws many other laws violated (wire fraud, identity theft, database rights, artistic rights), and standards breached (e.g. advertising standards). To complement this, it may be the case that a legal framework could be used to estimate the impact of information disorders based on the punitive measures associated with the laws and standards violated. Perhaps this could be used to inform deterrence strategies, such as publicizing punitive measures when laws are violated. Also, the real-world cases suggested some important adjustments for criterion 2. Nefarious actors clearly utilised a combination of techniques that spanned virtual and physical domains, and so only

focussing on the digital environment may be too restrictive. The complex interplay between digital and physical environments might be better captured as a hybrid approach to the distribution of WDID along with the dual impacts that would be estimated to occur (e.g., Danning, 2018; Stremlau et al., 2024, Van Raemdonck & Meyer, 2024).

Of the many challenges examining real-world cases presented, the most important is evidencing the impact of WDID. Regardless of how a risk-analysis framework articulates the impact, concrete examples are needed to capture the range of consequences associated with WDID. Extracting details to judge tangible consequences was difficult for reasons concerning the evolving nature of the WDID of the many cases examined, and the limited publicly available details. Along with this, the academic literature has considered WDID as situated in a broader range of cultural, societal, political and economic consequences (e.g., Danning, 2018; Lewis & Coaffee, 2024; Osman, 2025; Stremlau et al., 2024). Taken together, the concerns for an assessment of impact, means evidencing them, and reliably demonstrated the causal relationship between cases of WDID and their tangible and intangible consequences. While theories can make plausible cases for these types of associations, causal theorising, and sophisticated causal models are needed to explain (e.g., Grossmann et al., 2024; Shmueli, 2010) to then estimate the impacts of WDID.

A valid criticism that could be levelled at the selection of the real-world cases considered here is that they are not all prototypical examples of information disorders. The main motivation for exploring real-world cases is because the pollution of information ecospheres takes many forms, with a range of targets, and can cause damage in a variety of ways that should be considered. In fact, the limitation of the selection of the cases discussed here makes a strong case for having a comprehensive WDID incident database. The small selection of cases here can help to critically evaluate common assumptions, as well as point to ways in which risk concepts can be informed to better support a risk-analysis framework. Outside of this, a WDID incident database would be useful in its own right for researchers to explore patterns in the form WDID take, typical combinations that occur and their associated impacts. Data of this kind would help in theory development and empirical research on information disorders. In fact, to this point, much of the literature reviewed in the first section typically relies on idealised hypothetical examples, with few based on real-world case studies. There is no denying the need for tightly constructed scenarios and hypothetical cases to aid carefully controlled experiments, or for modelling, or testing the success of tools to detect WDID. However, a WDID incident database would clearly be of value to further test hypotheses and make predictions and to reduce current evidence gaps.

Conclusion

The present article discussed what, in principle, a risk-analysis framework estimating the impact of information disorders could look like. It also attempted a practical challenge by applying the proposed framework to a range of real instances of information disorders. What has been presented here is coarse on many levels and many limitations have been acknowledge throughout. The proposed risk-analysis framework is a starting point for informing future empirical, modelling, and theoretical work. The criteria and risk factors proposed need to be vigorously tested and robustly challenged to ensure that a future risk-analysis framework adequately addresses what could not be achieved presently. Using real-world cases is a useful exercise to explore what the future development of an incident database of WDID could look like. There are obvious benefits to having this not only for researchers in the risk analysis community, but for practitioners and decision-makers to make evidence-based decisions as to how to address the growing and alarming impacts of information disorders.

References

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) is misinformation a problem?. *Perspectives on Psychological Science*, *18*(6), 1436-1463.

Ahmad, W., Berg, R., & Kim, S. (2024). Combating Fake News with Digital Identity Verification. *URL: https://groups. csail. mit. edu/mac/classes/6.805/student-papers/fall17-papers/FakeNews. pdf [accessed 2024-09-17]*.

Ahmed, S., Masood, M., Bee, A. W. T., & Ichikawa, K. (2025). False failures, real distrust: the impact of an infrastructure failure deepfake on government trust. *Frontiers in Psychology*, *16*, 1574840.

Ahmed, M., & Pathan, A. S. K. (2020). False data injection attack (FDIA): an overview and new metrics for fair evaluation of its countermeasure. *Complex Adaptive Systems Modeling*, *8*(1), 4.

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social media+ society*, *9*(1), 20563051221150412.

Andersen, J., & Søe, S. O. (2020). Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news–the case of Facebook. *European Journal of Communication*, *35*(2), 126-139.

Andrei, M. (2024, July, 22). This cat is a highly-cited researcher — and it's a problem. ZME Science. https://www.zmescience.com/science/news-science/larry-richardson-cat-citations/

Arcuri, M. C., Gandolfi, G., & Russo, I. (2023). Does fake news impact stock returns? Evidence from US and EU stock markets. *Journal of Economics and Business*, *125*, 106130.

Arquam, M., Singh, A., & Sharma, R. (2021). A blockchain-based secured and trusted framework for information propagation on online social networks. *Social Network Analysis and Mining*, *11*(1), 49.

Asevameh, I. O., Dopamu, O. M., & Adesiyan, J. S. (2024). Election infrastructure security: a review of vulnerability and impact on the US economic reputation. *World Journal of Advanced Engineering Technology and Sciences*, *12*, 233-244.

Aven, T., & Thekdi, S. A. (2022). On how to characterize and confront misinformation in a risk context. *Journal of risk research*, *25*(11-12), 1272-1287.

Barker, K., Bessarabova, E., Radhakrishnan, S., González, A. D., Weber, M. S., Marquez, J. E. R., Vorobeychik, Y., & Jiang, J. N. (2025). Risk analysis of disinformation weaponized against critical networks. *Risk Analysis*, 1–9.

Barrett Peters, C. (2025, March, 12). MP questions NSW antisemitic laws after Dural caravan hoax. ABC News. https://www.abc.net.au/news/2025-03-12/nsw-dural-caravan-hoax-hate-speech-laws-antisemitism/105042380

Bennke, J. (2023). Media of Verification: An Epistemological Framework for Trust in a Digital Society. *communication+ 1*, *10*(1).

Berkowitz, E. (2021). *Dangerous ideas: A brief history of censorship in the West, from the ancients to fake News*. Saqi Books.

Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dutkiewicz, L., Krack, N., & Verdoliva, L. (2024). Generative AI and disinformation: recent advances, challenges, and opportunities.

Borkakoty, H., & Espinosa-Anke, L. (2024). Hoaxpedia: A unified Wikipedia Hoax articles dataset. *arXiv preprint arXiv:2405.02175*.

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, *48*(2), 139-166.

Burcă-Voicu, M. I., Cramarenco, R. E., & Dabija, D. C. (2025). Navigating the social media market: AI and the challenge of fake news dissemination in the business environment. *Oeconomia Copernicana*, *16*(1).

Butler, G. (2025, March, 10). Explosive-laden caravan plot was a hoax, say Australian police. BBC news. https://www.bbc.co.uk/news/articles/cly3zr8p46eo

Butler, S. (2025, May, 21). M&S expects cyber-attack to last into July and cost £300m in lost profits. Guardian UK. https://www.theguardian.com/business/2025/may/21/cyber-attack-cost-marks-and-spencer-lost-sales-company-results-reveal

Cappuccio, M. L., Sandis, C., & Wyatt, A. (2022). Online manipulation and agential risk. *The philosophy of online manipulation*, 72.

Caramancion, K.M. (2020). An exploration of disinformation as a cybersecurity threat. In Proceedings of the 2020 3rd IEEE International Conference on Information and Computer Technologies (ICICT), San Jose, CA, USA, 9–12 March 2020; pp. 440–444.

Caramancion, K. M., Li, Y., Dubois, E., & Jung, E. S. (2022). The missing case of disinformation from the cybersecurity risk continuum: A comparative assessment of disinformation with other cyber threats. *Data*, *7*(4), 49.

Chadwick, A., Vaccari, C., & Kaiser, J. (2025). The amplification of exaggerated and false news on social media: The roles of platform use, motivations, affect, and ideology. *American Behavioral Scientist*, *69*(2), 113-130.

Chalk, A. (2024, March, 4). Group that claimed it stole data from Epic admits it didn't happen, Epic says the whole thing was 'a scam. Yahoo!Tech. https://tech.yahoo.com/general/articles/group-claimed-stole-data-epic-225612419.html

Chan, K., & Sobhan, T. (2025, October, 29). Phony AI videos of Hurricane Melissa flood social media. Public Broadcasting Service. https://www.pbs.org/newshour/world/phony-ai-videos-of-hurricane-melissa-flood-social-media

Chang, C. C. (2022). Revisiting Disinformation Laws in the Age of Social Media. *Arizona Law of Emerging Technologies*, *6*, i.

Chechelashvili, M., Berikashvili, L., & Malania, E. (2023). Foreign interference in electoral processes as a factor of international politics: Mechanisms and counteraction. *Foreign Affairs*, (33), 52-62.

Cimili, Z. (2026, February, 12). NBC News. Actor takes legal action to stop Albania's government from using her image for 'AI minister'. Actor takes legal action to stop Albania's government from using her image for 'AI minister'

Cook, S. (2025, July, 8). "Unknown actor" using AI to impersonate Rubio, State Department cable shows. CBS news. https://www.cbsnews.com/news/marco-rubio-ai-impersonator-state-department-cable/

Danning, G. (2018). Did radio RTLM really contribute meaningfully to the Rwandan genocide?: Using qualitative information to improve causal inference from measures of media availability. Civil Wars, 20(4), 529–554. https://doi.org/10.1080/13698249.2018.1525677

Dhanji, K. (2025, October, 6). Deloitte to pay money back to Albanese government after using AI in $440,000 report. The Guardian. https://www.theguardian.com/australia-news/2025/oct/06/deloitte-to-pay-money-back-to-albanese-government-after-using-ai-in-440000-report

Doshi, J., Marino, J., Gan, S., Mager, D., Sprague, M., and Xia, M. (2023). Sleeper social bots: A new generation of AI disinformation bots are already a political threat. arXiv preprint arXiv:2408.12603. doi: 10.48550/arXiv.2408.12603

Douglas, K. M., & Sutton, R. M. (2023). What are conspiracy theories? A definitional approach to their correlates, consequences, and communication. *Annual review of psychology*, *74*(1), 271-298.

Eadon, Y. M., & Wood, S. E. (2025). Combating contamination and contagion: embodied and environmental metaphors of misinformation. *Convergence*, *31*(2), 500-520.

Echeverría, M., García Santamaría, S., & Hallin, D. C. (2025). *State-sponsored disinformation around the globe: How politicians deceive their citizens*. Taylor & Francis.

Ecker, U. K. H., Tay, L. Q., Roozenbeek, J., van der Linden, S., Cook, J., Oreskes, N., & Lewandowsky, S. (2025). Why misinformation must not be ignored. *American Psychologist, 80*(6), 867–878.

Eriksson Krutrök, M., & Lindgren, S. (2022). Social media amplification loops and false alarms: Towards a Sociotechnical understanding of misinformation during emergencies. *The Communication Review*, *25*(2), 81-95.

Fallis, D. (2014). The varieties of disinformation. In L. Floridi & Ph. Illari (Eds.), The philosophy of information quality (pp. 135–161). Springer.

Fawkes, J., & Gregory, A. (2001). Applying communication theories to the Internet. *Journal of communication management*, *5*(2), 109-124.

Ferrara, E. (2024). Charting the landscape of nefarious uses of generative artificial intelligence for online election interference. *arXiv preprint arXiv:2406.01862*.

Fenton, N., & Neil, M. (2013). *Risk assessment and decision analysis with Bayesian networks*. Boca Raton FL: Crc Press, Taylor & Francis Group.

Fleming, C., & O'Carroll, J. (2010). The art of the hoax. *parallax*, *16*(4), 45-59.

Gabriel, S., Lyu., L, Siderius, J., Ghassemi, M., Andreas, J., & Ozdaglar, A. (2024). 2024. "Generative AI in the Era of 'Alternative Facts.'" *An MIT Exploration of Generative AI*, March. https://doi.org/10.21428/e4baedd9.82175d26.

García-Ull, F. J., Broseta-Dupré, B., & Lamirán-Palomares, J. M. (2025). Analyzing X mentions to uncover micro-clusters interfering

in political campaigns: a case study during the Spanish elections. *Frontiers in Communication*, *10*, 1545634.

Gavurova, B., Moravec, V., Hynek, N., Miovsky, M., Polishchuk, V., Gabrhelik, R., Bartak, M., Petruzelka, B., & Stastna, L. (2024). The impact of digital disinformation on quality of life: a fuzzy model assessment. *Technological and Economic Development of Economy*, *30*(4), 1120-1145.

Gelfert, A. (2018). Fake news: A definition. *Informal logic*, *38*(1), 84-117.

Granados Samayoa, J. A., & Albarracín, D. (2025). Understanding belief-behavior correspondence: Beliefs and belief-to-behavior inferences. *Psychological Inquiry*, *36*(1), 1-22.

Greig, J. (2025, April, 25). Largest telecom in Africa warns of cyber incident exposing customer data. The Record. https://therecord.media/largest-african-telecom-warns-of-data-exposure

Grossmann, I., Varnum, M. E., Hutcherson, C. A., & Mandel, D. R. (2024). When expert predictions fail. Trends in Cognitive Sciences, 28(2), 113–123. https://doi.org/10.1016/j.tics.2023.10.005

Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2022). Social bots and the spread of disinformation in social media: the challenges of artificial intelligence. *British Journal of Management*, *33*(3), 1238-1253.

Hals, T. (2025, July, 17). Meta investors, Zuckerberg reach settlement to end $8 billion trial over Facebook privacy violations. Reuters. https://www.reuters.com/sustainability/boards-policy-regulation/meta-investors-zuckerberg-reach-settlement-end-8-billion-trial-over-facebook-2025-07-17/

Hameleers, M. (2023). Disinformation as a context-bound phenomenon: toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*, *33*(1), 1-10.

Harashima, T. (2023). An Economic Theory of Disinformation. *Theoretical and Practical Research in Economic Fields (TPREF)*, *14*(27), 16-28.

Henderson, E. H. (1943). Toward a definition of propaganda. *The Journal of Social Psychology*, *18*(1), 71-87.

Ibrahim, H., Liu, F., Zaki, Y., & Rahwan, T. (2025). Citation manipulation through citation mills and pre-print servers. *Scientific reports*, *15*(1), 5480.

International Organization for Standardization (2019). Risk Management – Risk assessment

techniques. https://www.iso.org/standard/72140.html

Jamalzadeh, S., Mettenbrink, L., Barker, K., González, A. D., Radhakrishnan, S., Johansson, J., & Bessarabova, E. (2024). Weaponized disinformation spread and its impact on multi-commodity critical infrastructure networks. *Reliability Engineering & System Safety*, *243*, 109819.

Jurgenson, N. (2011). Digital dualism versus augmented reality. *The Society Pages*, *24*.

Kausar, F., Deo, S., Hussain, S., & Ul Haque, Z. (2024). Federated Deep Learning Model for False Data Injection Attack Detection in Cyber Physical Power Systems. *Energies*, *17*(21), 5337.

Khameneh, R. T., Barker, K., & Ramirez-Marquez, J. E. (2025). A hybrid machine learning and simulation framework for modeling and understanding disinformation-induced disruptions in public transit systems. *Reliability Engineering & System Safety*, *255*, 110656.

Kissin, E. (2025, October, 6). Deloitte issues refund for error-ridden Australian government report that used AI. Financial Times. https://www.ft.com/content/934cc94b-32c4-497e-9718-d87d6a7835ca

Kleinman, Z. (2025, July, 4). AI claims and a hoax spokesman: Viral band confuses the world of music. BBC news. https://www.bbc.co.uk/news/articles/cp8mjnn7eqno

Krishna, T. (2024, April, 10). NSE raises the red flag! Cautions against fake videos of MD recommending stocks. Financial Express. https://www.financialexpress.com/market/nse-raises-the-red-flag-cautions-against-fake-videos-of-md-recommending-stocks-3452439/

Kyriakidou, M., Morani, M., Cushion, S., & Hughes, C. (2023). Audience understandings of disinformation: navigating news media through a prism of pragmatic scepticism. *Journalism*, *24*(11), 2379-2396.

Kumar, S., Sai, S., Chamola, V., Gaur, A., Agarwal, C., Huang, K., & Hussain, A. (2025). Peeping into the Future: Understanding and Combating Generative AI-Based Fake News. *Cognitive Computation*, *17*(3), 1-33.

Kwon, K. H., Lee, M. H., Pil Han, S., & Park, S. (2024). Fake thumbs in play: A large-scale exploration of false amplification and false diminution in online news comment spaces. *New Media & Society*, *26*(6), 3252-3272.

Leng, C., & Ho-him, C. (2024, May, 17). Arup lost $25mn in Hong Kong deepfake video conference scam. Financial Times. https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea

Levak, T. (2020). Disinformation in the new media system–Characteristics, forms, reasons for its dissemination and potential means of tackling the issue. *Medijska istraživanja: znanstveno-stručni časopis za novinarstvo i medije*, *26*(2), 29-58.

Lewis, T., & Coaffee, J. (2024). Atmosphere, imminence and the Manchester arena inquiry: On the affective modalities of becoming situationally aware to urban terrorism. Critical Studies on Security, 13(1), 38–57. https://doi.org/10.1080/21624887.2024.2385185

Li, T. (2025). Privacy and Disinformation. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5318708

Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading voters using human-artificial intelligence dialogues. Nature. 648, 394–401.

Lindaas, O., & Bakken, B. (2025). The epistemology of risk. Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage & Marja Ylönen. Proceedings of the 35th European Safety and Reliability and the 33th Society for Risk Analysis Europe Conference, Stavanager, Norway.

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., ... & Peng, W. (2024). A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Liu, Z., Liu, M., Wang, Q., & Tang, Y. (2024). False Data Injection Attacks on Data-Driven Algorithms in Smart Grids Utilizing Distributed Power Supplies. *Engineering*.

Maddox, A. (2017). Beyond digital dualism: Modeling digital community. *Digital sociologies*, *1*.

Maimann, K. (2025, July, 5). How a Canadian's AI hoax duped the media and propelled a 'band' to streaming success. CBC news. https://www.cbc.ca/news/entertainment/ai-band-hoax-velvet-sundown-1.7575874

Májovský, M., Černý, M., Kasal, M., Komarc, M., & Netuka, D. (2023). Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *Journal of medical Internet research*, *25*(1), e46924.

Markov, Č., & Min, Y. (2022). Understanding the public's animosity toward news media: Cynicism and distrust as related but distinct negative media perceptions. *Journalism & Mass Communication Quarterly*, *99*(4), 1099-1125.

Mascelliono, A. (2024, November, 18). Swiss Cyber Agency Warns of QR Code Malware in Mail Scam. Infosecurity Magazine. https://www.infosecurity-magazine.com/news/swiss-cyberagency-qr-code-mail-scam/

McLoughlin, K. L., & Brady, W. J. (2024). Human-algorithm interactions help explain the spread of misinformation. *Current opinion in psychology*, *56*, 101770.

Mendelson, A. (202k, May, 20). Beauty queen sold 'fibre' gummies laced with laxatives. Yahoo!news.

https://www.yahoo.com/news/beauty-queen-sold-fibre-gummies-092559192.html

Metzler, H., & Garcia, D. (2024). Social drivers and algorithmic mechanisms on digital media. *Perspectives on Psychological Science*, *19*(5), 735-748.

Milmo, D. (2024, May, 17). UK engineering firm Arup falls victim to £20m deepfake scam. The Guardian. https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video

Mitrovic, Z. (2025, April, 25). Data Poisoning: The Silent AI Killer Reshaping Ransomware Threats in 2025. Linkedin. Data Poisoning: The Silent AI Killer Reshaping Ransomware Threats in 2025

MTN Group (2025, April, 24). MTN cybersecurity incident, but critical infrastructure secure. MTN Media release. https://www.mtn.com/mtn-cybersecurity-incident-but-critical-infrastructure-secure/

Mulcahy, R., Barnes, R., de Villiers Scheepers, R., Kay, S., & List, E. (2024). Going viral: Sharing of misinformation by social media influencers. *Australasian Marketing Journal*, 14413582241273987.

Musafer, S., & McMahon, L. (2025, May, 1). What can I buy online at M&S since the hack? BBC news. https://www.bbc.co.uk/news/articles/c0el31nqnpvo

National Cyber security center NCSC. (2024, November, 11). Caution: Fake letters on behalf of MeteoSwiss – Instead of a 'Severe Weather Warning App', malware is downloaded. https://www.ncsc.admin.ch/ncsc/en/home/aktuell/im-fokus/2024/2024-meteosuisse.html

Neil, M., Fenton, N., Osman, M., & Lagnado, D. (2021). Causality, the critical but often ignored component guiding us through a world of uncertainties in risk assessment. *Journal of Risk Research*, *24*(5), 617-621.

Ó Fathaigh, R., Helberger, N., & Appelman, N. (2021). The perils of legally defining disinformation. *Internet policy review*, *10*(4), 2022-40.

Oksanen, A., Celuch, M., Oksa, R., & Savolainen, I. (2024). Online communities come with real-world consequences for individuals and societies. *Communications Psychology*, *2*(1), 71.

Olivares-Delgado, F., Benlloch-Osuna, M., Rodríguez-Valero, D., & Breva-Franch, E. (2022). Corporate disinformation: Concept and typology of forms of corporate disinformation. In *International Conference on Design and Digital Communication* (pp. 536-550). Cham: Springer Nature Switzerland.

Osman, M. (2024). Public evaluations of misinformation and motives for sharing it. *Journalism and Media*, *5*(2), 766-786.

Osman, M. (2025). Evidencing the Impact of Misinformed and Disinformed Beliefs on Individual and Group Behaviors. *Psychological Inquiry*, *36*(1), 49-56.

Osman, M., Adams, Z., Meder, B., Bechlivanidis, C., Verduga, O., & Strong, C. (2022). People's understanding of the concept of misinformation. *Journal of Risk Research*, *25*(10), 1239-1258.

Pamment, J. (2022). *A capability definition and assessment framework for countering disinformation, information influence, and foreign interference*. NATO Strategic Communications Centre of Excellence.

Paul, O., & Yesmin, S. (2024). Rethinking the Misinformation with its Detrimental Impact on Lives: A Qualitative Approach. *Science & Technology Libraries*, *43*(3), 251-266.

Pérez-Escolar, M., Lilleker, D., & Tapia-Frade, A. (2023). A systematic literature review of the phenomenon of disinformation and misinformation. *Media and communication*, *11*(2), 76-87.

Petratos, P. N., & Faccia, A. (2023). Fake news, misinformation, disinformation and supply chain risks and disruptions: risk management and resilience using blockchain. *Annals of Operations Research*, *327*(2), 735-762.

Peukert, C., & Windisch, M. (2025). The economics of copyright in the digital age. *Journal of Economic Surveys*, *39*(3), 877-903.

Pierce, G. L., Holland, C. C., Cleary, P. F., & Rabrenovic, G. (2022). The opportunity costs of the politics of division and disinformation in the context of the twenty-first century security deficit. *SN Social Sciences*, *2*(11), 241.

Plikynas, D., Rizgelienė, I., & Korvel, G. (2025). Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact through Machine Learning. *IEEE Access*.

Polčák, R., & Kasl, F. (2021). Proportionate Forensics of Disinformation and Manipulation. In *Challenging Online Propaganda and Disinformation in the 21st Century* (pp. 167-193). Cham: Springer International Publishing.

Pooranian, Z., Conti, M., Haddadi, H., & Tafazolli, R. (2021). Online advertising security: Issues, taxonomy, and future directions. *IEEE Communications Surveys & Tutorials*, *23*(4), 2494-2524.

Porter, E., Shaw, A., & Spiro, E. S. (2025). Misinformation research continues to be urgent science. *Science Advances*, *11*(22), eady9839.

Rathje, S., & Van Bavel, J. J. (2025). The psychology of virality. *Trends in Cognitive Sciences*.

Rickards, J. (2024). *MoneyGPT: AI and the Threat to the Global Economy*. Penguin.

Rodríguez-Ferrándiz, R. (2023). An overview of the fake news phenomenon: From untruth-driven to post-truth-driven approaches. *Media and Communication*, *11*(2), 15-29.

Roozenbeek, J., & Van der Linden, S. (2024). The psychology of misinformation. Cambridge University Press.

Ruohonen, J. (2024). A comparative study of online disinformation and offline protests. *SN Social Sciences*, *4*(12), 232.

Ruiz, C. D. (2025). *Market-oriented disinformation research: Digital advertising, disinformation and fake news on social media* (p. 232). Taylor & Francis.

Shubham, S., Mehta, A., & Malhotra, M. V. (2025). Exploring Security Strategies: Safeguarding Networks From Cyber Attacks With Advanced Cyber Security Techniques. *International Journal of Environmental Sciences*, *11*(9s), 1151-1159.

Schünemann, W. J. (2022). A threat to democracies?: An overview of theoretical approaches and empirical measurements for studying the effects of disinformation. *Cyber security politics*, 32-47.

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, *116*(16), 7662-7669.

Schroeder, D., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N., Garcia, D., Goldenberg, A., Kyrychenko, Y., Leyton-Brown, K., Lutz, N., Marcus, G., Menczer, F., Pennycook, G., Rand, D., Ressa, M., Schweitzer, F., Song, D., Summerfield, C., Tang, A., Van Bavel, J., van der Linden., S., & Kunst, J. (2026). How malicious AI swarms can threaten democracy. *Science*, *391*, 354-357.

Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3), 289–-310.

Shutler, A. (2024, March, 4). Hacker group admits Epic Games breach was scam to catch other criminals. NME. https://www.nme.com/news/gaming-news/hacker-group-admits-epic-games-breach-was-scam-to-catch-other-criminals-3596497

Sobhan, T., & Chan, K. (2025, October, 30). Phony AI-generated videos of Hurricane Melissa flood social media sites. The Boston Globe. https://www.bostonglobe.com/2025/10/30/business/phony-ai-videos-hurricane-melissa/

Søe, S. O. (2021). A unified account of information, misinformation, and disinformation. *Synthese*, *198*(6), 5929-5949.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & language*, *25*(4), 359-393.

Stremlau, N., McGeer, C., & Straub, M. (2024). Deciphering digital hate: Assessing the evidence between online speech and offline violence in Africa. Global Media Journal—German Edition, 13(2) https://doi.org/10.60678/gmj-de.v13i2.280

Trammell III, T. I. (2020). *Fake News Risk: Modeling Management Decisions to Combat Disinformation*. Stanford University.

Utami, P. (2018). Hoax in modern politics: The meaning of hoax in Indonesian politics and democracy. *Jurnal Ilmu Sosial dan Ilmu Politik*, *22*(2), 85-97.

Van Hoboken, J., Appelman, N., Ó Fathaigh, R., Leerssen, P., McGonagle, T., van Eijk, N., & Helberger, N. (2019). The legal framework on the dissemination of disinformation through Internet services and the regulation of political advertising [Report]. Dutch Ministry of Interioor and Kingdom Relations. https://www.ivir.nl/publicaties/download/Report_Disinformation_Dec2019-1.pdf

van Hoboken, J., & Fathaigh, R. Ó. (2021). Regulating Disinformation in Europe: Implications for Speech and Privacy. *UC Irvine Journal of International, Transnational, and Comparative Law.*, *6*, 9-36.

van der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2025). Using psychological science to understand and fight health misinformation: An APA consensus statement. *American Psychologist.* Advance online publication. https://doi.org/10.1037/amp0001598

Van Prooijen, J. W., & Van Vugt, M. (2018). Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on psychological science*, *13*(6), 770-788.

Van Raemdonck, N., & Meyer, T. (2024). Why disinformation is here to stay. A socio-technical analysis of disinformation as a hybrid threat. In *Addressing hybrid threats* (pp. 57-83). Edward Elgar Publishing.

Varela da Costa, J., Dongo, D. F., & Mira da Silva, M. (2025). Using MCDA to select countermeasures against fake news. *Journal of Information, Communication and Ethics in Society*, *23*(1), 54-103.

Vasist, P. N., Chatterjee, D., & Krishnan, S. (2024). The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information Systems Frontiers*, *26*(2), 663-688.

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

Walton, D. (1997). What is propaganda, and what exactly is wrong with it. *Public Affairs Quarterly*, *11*(4), 383-413.

Wells, G., Romhanyi, A., Reitman, J. G., Gardner, R., Squire, K., & Steinkuehler, C. (2024). Right-wing extremism in mainstream games: A review of the literature. Games and Culture, 19(4), 469–492.

White, R. (2025, May, 28). A weaponized AI chatbot is flooding city councils with climate misinformation. https://www.nationalobserver.com/2025/05/28/investigations/weaponized-ai-chatbot-city-councils-climate-misinformation.

Wilcox, C. (2024, August, 1). ScienceAdviser: Meet Larry, the world's most highly cited cat. Science. https://www.science.org/content/article/science-adviser-meet-larry-the-worlds-most-highly-cited-cat

Wong, T. (2025, May, 20). Vietnamese beauty queen arrested for fraud over fibre gummies. BBC news. https://www.bbc.co.uk/news/articles/c2e33dvpgxro

Whyte, C. (2018). Crossing the digital divide: Monism, dualism and the reason collective action is critical for cyber theory production. *Politics and Governance*, *6*(2), 73-82.

Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Yee, A. K. (2023). Information deprivation and democratic engagement. *Philosophy of Science*, *90*(5), 1110-1119.

Ylönen, M., & Aven, T. (2023). A framework for understanding risk based on the concepts of ontology and epistemology. *Journal of risk research*, *26*(6), 581-593.

Young, K. (2017). *Bunk: The rise of hoaxes, humbug, plagiarists, phonies, post-facts, and fake news*. Graywolf Press.

Yousif, N. (2025, July, 8). Imposter used AI to pose as Marco Rubio and contact foreign ministers. BBC news. https://www.bbc.co.uk/news/articles/crrqkyyjewno

Yousif, N., & Jamali, L. (2025, July, 17). Meta investors settle $8bn lawsuit with Zuckerberg over Facebook privacy. BBC news. https://www.bbc.co.uk/news/articles/cx2jmledvr3o

Zhou, K., Šćepanović, S., & Quercia, D. (2024). Characterizing Fake news targeting corporations. In *Proceedings of the International AAAI Conference on Web and Social Media*, 18, 1818-1832.