costaid

Coping Strategies Against Information Disorder

Guidelines for first-liners

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Authors:

Eliane Smits van Waesberghe & Tim Paulusse – Verwey-Jonker Instituut (Main Editors)

Leen D'Haenens & Joyce Vissenberg – KU Leuven

Tzvetalina Genova – International Management Institute

Wolfgang Eisenreich – Wissenschaftsinitiative Niederösterreich

Sonja Bercko Eisenreich – Integra Institute

Alenka Valjašková – QUALED

Pantelis Balaouras – Connexions

Declaration on copyright:

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

You are free to:

- share — copy and redistribute the material in any medium or format
- adapt — remix, transform, and build upon the material

under the following terms:

- Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- NonCommercial — You may not use the material for commercial purposes.
- ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original

costaid
Coping Strategies Against Information Disorder

Chapter 3
Technology & Tools

Target group

These guidelines are targeted towards so-called “first-liners”. “First-liners” is an overarching term for all people in direct contact with people who are vulnerable to information disorder, focused on groups in vocational education. Examples of people who fall under the umbrella term are: educators, teachers, trainers, youth counsellors and advisors, social workers and youth workers. This is a non-exhaustive list, however. The scope of this project also includes other people working in the educational, social or health care sector.

3.1 Introduction

The concept of misinformation (fabricated news) has been discussed in previous chapters. News is typically disseminated by professional news providers, including public service media, commercial news media, independent professional journalism, or amateur users (in the case of social media platforms). News is accessible and available in various formats:

- **Digital News**: News distributed through internet-based channels in digital media format (text, pictures, audio, video).
- **Print Media**: Newspapers and magazines containing text and image-based content.
- **Broadcasting**: TV and radio with video and audio-based content.

In this section, our focus primarily revolves around the first format, digital news. However, the discussions presented may also be applicable to other formats that utilise video, audio, and pictures.

Digital news is predominantly accessed through news platforms operated by professional news providers. Users may access these news platforms either for free or through subscription services. These platforms offer web feeds within popular browsers and news feeds on social media platforms, allowing users to receive personalised news updates.

A news feed is a web page or screen that frequently updates to display the latest news or information. Personalised news feeds are services integrated into web browsers (web feeds) and social media platforms that deliver news to users based on their personal preferences. These preferences are determined by various factors such as subscribing to web feed channels, visiting specific web pages, and more.

In social media platforms, users also receive shared media from other users, which may include instances of misinformation.

Personalised news appears on a user's browser or social media platform through services and social media algorithms. Users may not be fully aware of how these algorithms operate, as it remains unclear whether the selection of news is based solely on user preferences or other criteria. For example, social media platforms present news based on the news providers a user follows, the reading habits of their friends, or recently clicked articles, which may form a type of preference. Consequently, a list of news is filtered, meaning that not all news is displayed but rather those that the algorithm deems most interesting to the user. Many argue that this creates an information bubble, wherein news is selected by an algorithm and may not include news that a user is interested in, but the algorithm fails to include. Therefore, it is important not to solely rely on personal news feeds but rather visit professional news provider platforms that are trusted. It is always important to critically evaluate the credibility and accuracy of news, regardless of the format or delivery method.

3.2 Search engines & algorithms

Many websites on the internet aim to keep their users engaged and maximise their time spent on the platform. This is especially true for social media platforms, which employ various strategies to enhance the user experience. One such strategy is showing users content that aligns with their interests. However, websites cannot read minds to know users' preferences. To address this, algorithms are used to analyse user data and deliver personalised content.

Algorithms are complex formulas that observe, measure, and calculate an individual's content preferences. This can include factors such as the user's watch time on specific types of videos, the time spent on a particular post, or engagement actions like leaving likes or comments. By analysing this data, algorithms determine the type of content that keeps users engaged for longer periods. This information is then used to curate and recommend similar content to the user.

While this approach may seem logical and harmless, there are potential downsides to algorithmic use. The algorithm's preference for content the user finds interesting can create filter bubbles, where users are exposed only to specific viewpoints. Filter bubbles restrict the diversity of content, potentially leading to echo chambers or reinforcing existing ones.

In addition to filter bubbles, algorithms can amplify the extremity of content by recommending increasingly niche, fringe, and extreme posts. The goal is to keep users engaged, but this can lead to users being funneled into online spaces devoid of differing viewpoints, resulting in what is called a "rabbit hole".

Filter bubbles and rabbit holes expose users to radical content and the communities associated with it. These online spaces provide fertile ground for the development, growth, distortion, and propagation of misinformation and disinformation.

As users progress through the rabbit hole, extreme talking points and false information become normalised. This normalisation further facilitates the spread and acceptance of misinformation and disinformation from radical sources.

By understanding the impact of algorithms on user experiences, we can better comprehend the risks associated with filter bubbles, echo chambers, and rabbit holes. This knowledge is essential for navigating the online landscape and addressing the challenges posed by information disorder.

3.3 Online strategies of disinformation & extremist organisations

In the previous sections, we explored how information disorder is created and perpetuated within echo chambers and filter bubbles. But how does false information reach people outside of those spaces?

The most common methods

Mis- and disinformation can be spread through countless forms of communication. However, they are most typically spread through various channels, such as social media platforms, websites, email, and word of mouth. The most common methods by which information disorder is spread are closely related to the seven categories of problematic content which are discussed in *Chapter 1: Understanding 'fake news'*:

- **Satire or Parody**: Some mis- and disinformation is created for entertainment purposes or satire but can be misconstrued as genuine news. Satirical websites or social media accounts may publish humorous or exaggerated stories, but readers who are unaware of their satirical nature can mistake them for factual information.
- **Clickbait**: articles which include mis- or disinformation often employ sensational or misleading headlines to grab attention and generate more clicks or views. They aim to exploit people's curiosity or emotions to drive traffic to a website and generate revenue through advertising.
- **Misrepresentation**: This involves distorting or misrepresenting actual news by selectively presenting facts or omitting crucial information. It can involve taking statements out of context, altering images or videos, or twisting the meaning of a story to fit a particular narrative.
- **Impersonation**: mis- and disinformation can also involve impersonating reputable news sources or public figures to lend credibility to false information. This can be done through creating fake websites or social media accounts that mimic legitimate sources, fooling readers into believing the information is trustworthy.
- **Political manipulation**: Information disorder is sometimes created or spread with the intention of influencing public opinion or elections. This can involve spreading false information about political candidates, manipulating public sentiment, or exploiting existing biases and divisions within society.
- **Fabrication**: mis- and disinformation can be entirely fabricated, with no basis in reality. It involves creating false stories, quotes, or events to mislead readers or viewers.
- **Amplification through social media**: Social media platforms play a significant role in the spread of mis- and disinformation. False stories can quickly

go viral as users share and repost them, often without verifying the accuracy of the information. The algorithms used by social media platforms can also contribute to the amplification by promoting content based on engagement rather than accuracy.

There are also other common methods used to spread mis- and disinformation which should be discussed in more detail: trolling, doxxing and mainstreaming. This is done in the next two subchapters.

Trolling & doxxing

One frequently employed strategy is trolling. Trolling is defined as the deliberate use of impoliteness, aggression, deception, and manipulation in online communication to provoke conflict or amusement. Trolls instigate online conflicts by deceiving, manipulating, or being aggressive. They derail conversations for their own amusement, essentially engaging in digital bullying.

On a small scale, trolling may seem relatively harmless, appearing as mere annoyance. However, when organised groups of trolls share a specific goal, this annoyance can rapidly transform into a disinformation epidemic. An example of this is Russia's utilisation of social media trolls as a 'weapon'. Russia employed a large network of trolls to globally disseminate disinformation in multiple languages, aiming to control the online discourse surrounding Russia. These trolls not only spread false information but also targeted social media users with posts that deviated from the narrative they were organised to promote. Consequently, many social media users refrained from discussing Russia, effectively allowing the trolls to control the narrative with their misinformation.

Doxxing, another form of internet bullying, involves revealing personal information or identities of individuals online without their consent. While this tactic does not specifically involve the spread of disinformation, it is another strategy that internet trolls employ to control the narrative on a particular topic, similar to trolling. Doxxing can be used to intimidate social media users, suppressing their willingness to post content that goes against the troll's preferred narrative.

Understanding the impact of trolling and doxxing is crucial in recognising the various tactics employed to manipulate and control online narratives. These strategies not only contribute to the spread of false information but also pose challenges in fostering an open and informed digital environment.

Mainstreaming

An important strategy to disseminate disinformation and extremist content is through normalising or “mainstreaming” it. Exposure plays a crucial role in this process. The exposure to mis- and disinformation can lead to persistent misconceptions of people surrounding the particular topics, which normalises the false ideas in their minds. This exposure can happen in various forms.

Influencers

One common form of exposure is person-to-person dissemination, where individuals share information with others. This can occur through personal interactions or on a larger scale with influencers on social media. Influencers, who have a significant reach across different groups, can unknowingly or deliberately share false information, impacting a large number of individuals. Such widespread exposure leads to the normalisation of misinformation among diverse audiences.

Irony, satire & memes

Extremist individuals and organisations often incorporate humor, satire, and irony to spread their ideas.

Satire can be a powerful tool to challenge oppressive ideologies, shift narratives, or normalise niche views within the mainstream. In the realm of misinformation, satire is employed on a spectrum. Parody websites like The Onion or De Speld publish non-factual content for humor purposes, without intending to deceive the public. However, certain individuals and groups utilise satire and irony with malicious intent to discredit mainstream journalism, science, or promote extremist ideas and disinformation. By leveraging satire and humor, such content becomes more accessible and acceptable in political discourse, exposing it to a wider audience.

Extremist content often appeals to young people as a form of entertainment or sensation-seeking. Young individuals, driven by a search for meaning, tend to gravitate towards intense and novel experiences, making them more susceptible to extremist ideas and the associated disinformation.

Memes, which are widely shared pieces of humorous cultural content, serve as another avenue for spreading extremist ideologies. Memes come in various formats, including pictures, videos, audio clips, emojis, and symbols. While memes themselves are not inherently harmful, extremists use them to normalise their ideas. The playful nature of memes allows extremists to disguise, debunk, or deny the harmfulness of their messaging. This "edgy" or provocative content becomes more acceptable, and when confronted with accusations of sexism, racism, or xenophobia, the creators can easily dismiss it as "just a joke". This blurring of boundaries between playful mischief and problematic content creates ambiguity, making it challenging to discern innocent jokes from extremist messaging. Pepe the Frog, an internet cartoon character which had been initially created to be a harmless

joke, was appropriated by online white supremacists. This caused confusion with internet users, because extremist iterations of this meme found themselves mixed in with the harmless ones. The normalisation of extremist content occurs as more individuals are exposed to these messages, blurring the lines between what is acceptable and what is not.

3.4 Manipulated content

From a technical perspective, all information or "news" in media is a combination of text, picture, audio, and video. However, the concern lies in determining whether the information is authentic or not. It is worth noting that misinformation may utilise genuine pictures but manipulate the story, distorting the actual facts.

In the past, it was commonly understood that anyone could write a text, while pictures, audio, and video were assumed to be more or less authentic, requiring professional skills for modifications. However, with recent technological advancements, even pictures, audio, and video can be subject to manipulation. This can be achieved by professionals or through applications utilising artificial intelligence systems, such as deepfake technology. Consequently, it is necessary to distinguish whether an audio or video has been genuinely captured by a microphone or video camera, or if it is a result of expert editing or artificial intelligence (AI) systems (generative AI and synthetic media: voice clones, **deepfake** videos). Additionally, it should be technically feasible to identify the original source, producer, or publisher of a picture, audio, or video resource. This is because resources can be shared, copied, or redistributed numerous times across the World Wide Web and social media. Therefore, for regular users, it can be challenging to identify the original source and producer, even if they suspect that the news may be misinformation.

Empowering users to distinguish between real and fabricated news requires several steps. More information on this can be found in *Chapter 2: Actions & Skills*. It is encouraged to read this chapter in order to learn the intricate details behind recognising false information. However, here is a short, very simplified summary:

- **Step 1: Foster User Awareness:** Users need to be aware that news can be fabricated. Conducting awareness activities is crucial to inform users about what constitutes fabricated news and how they can protect themselves from its consequences.
- **Step 2: Verify Publisher Reliability**: Increased awareness of fabricated news prompts users to question the reliability of news sources and publishers. It is essential to consider the media format, whether it is a TV channel, journal, newspaper (online or printed), or a social media platform. Media channels that allow easy sharing or redistribution of news may be less reliable. Conversely, media channels that facilitate source identification and verification tend to be more reliable.

Regarding news distributed over the internet, service providers such as News Feeds and Social Media Networks should leverage emerging technologies to verify source reliability and track the original publisher and source. Blockchain technology is one such technology that can facilitate these efforts.

By following these steps and leveraging technology, users can be empowered to navigate the digital landscape, distinguish real news from misinformation, and make informed decisions about the information they encounter.

Understanding deepfakes: synthetic media manipulation

Deepfakes, as defined by the Cambridge Dictionary, are "*video or sound recordings that replace someone's face or voice with that of someone else, in a way that appears real*".

In the article "Deepfake explained" from 2020, the writer Meredith Somers mentions that "*(a) deepfake refers to a specific kind of synthetic media where a person in an image or video is swapped with another person's likeness*". Furthermore, it is explained that "*the term 'deepfake' was first coined in late 2017 by a Reddit user of the same name. This user created a space on the online news and aggregation site, where they shared pornographic videos that used open-source face-swapping technology."*

Deepfakes have found applications in various sectors and have been used for different purposes. Some notable examples include:

- **Blackmail:** Deepfakes can be used to generate false incriminating material, potentially leading to blackmail. Moreover, as it becomes increasingly difficult to distinguish deepfakes from genuine content, victims of actual blackmail can claim that the evidence is fake, granting them plausible deniability.
- **Pornography:** Deepfake pornography has gained significant prominence on the internet. A report by Dutch cybersecurity startup Deeptrace estimated that approximately 96% of all online deepfakes were pornographic.
- **Politics:** Deepfakes have been utilised to misrepresent well-known politicians in videos, spreading disinformation and manipulating public perception. Examples include deepfakes featuring Barack Obama, Donald Trump, Volodymyr Zelenskyy, and Vladimir Putin.
- **Acting/Films:** Speculation exists regarding the use of deepfakes for creating digital actors in future films. While digitally constructed or altered humans have been featured in films before, deepfakes could contribute to new advancements in this domain.
- **Social media:** Deepfakes have been utilised by users on various social media platforms. Individuals replace faces in popular film or series scenes with their own, creating personalised videos. Platforms like Facebook have taken measures to detect and flag deepfakes as fake, reducing their priority in users' feeds.

Detecting and mitigating deepfakes: technological approaches

Researchers are actively exploring methods to detect and address the issue of deepfake audio and video. Various approaches are being pursued:

- **Algorithmic Detection:** One approach involves developing algorithms that can identify manipulated content. These algorithms analyse various visual and auditory cues to detect inconsistencies or anomalies that indicate the presence of a deepfake. By leveraging machine learning and artificial intelligence techniques, these algorithms can improve their detection capabilities over time.
- **Blockchain Technology:** Another technique proposes utilising blockchain technology to verify the source of media. The blockchain is a digital ledger that records transactions across a network of computers in a secure, transparent, and tamper-resistant way. It uses decentralisation and cryptography to ensure trust without the need for a central authority. In this scenario, videos would need to undergo verification through a blockchain ledger before being displayed on social media platforms. By ensuring that only videos from trusted sources are approved, the spread of potentially harmful deepfake media could be reduced.
- **Digital Signatures:** Some suggest digitally signing all videos and imagery captured by cameras, including smartphone cameras, as a means to combat deepfakes. This would involve assigning unique digital signatures to each piece of media, enabling the tracing of every photograph or video back to its original owner. While this approach can aid in tracking the origin of content, there are concerns regarding potential misuse, such as targeting dissidents or violating privacy.

3.5 References

Aro, J. (2016). The Cyberspace War: Propaganda and Trolling as Warfare Tools. *European View*, *15*(1), 121–132. https://doi.org/10.1007/s12290-016-0395-5

Cambridge English Dictionary: Meanings & Definitions. (2023). https://dictionary.cambridge.org/dictionary/english

Daniels, J. (2018). The Algorithmic Rise of the "Alt-Right." *Contexts*, *17*(1), 60–65. https://doi.org/10.1177/1536504218766547

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: a framework and research agenda. *Annals of the International Communication Association*, *43*(2), 97–116. https://doi.org/10.1080/23808985.2019.1602782

Greene. (2019). "Deplorable" Satire: Alt-Right Memes, White Genocide Tweets, and Redpilling Normies. *Studies in American Humor*, *5*(1), 31–69. https://doi.org/10.5325/studamerhumor.5.1.0031

Hardaker, C. (2013). "Uh. . . . not to be nitpicky,,,,,but. . .the past tense of drag is dragged, not drug." *Journal of Language Aggression and Conflict*, *1*(1), 58–86. https://doi.org/10.1075/jlac.1.1.04har

IED. (2022, August 23). *How Do Social Media Algorithms Work*. Institute of Entrepreneurship Development. https://ied.eu/blog/technology-blog/how-do-social-media-algorithms-work/

Johnson, D., & Johnson, A. (2023, June 15). What are deepfakes? How fake AI-powered audio and video warps our perception of reality. *Business Insider*. https://www.businessinsider.com/guides/tech/what-is-deepfake?international=true&r=US&IR=T

Levy, G., & Razin, R. (2019). Echo Chambers and Their Effects on Economic and Political Outcomes. *Annual Review of Economics*, *11*, 303–328. https://doi.org/10.1146/annurev-economics-080218-030343

Lewis, B., & Marwick, A. E. (2017). Media Manipulation and Disinformation Online. *Data & Society Research Institute*. https://www.posiel.com/wp-content/uploads/2016/08/Media-Manipulation-and-Disinformation-Online-1.pdf

McNealy, J. (2015). Readers react negatively to disclosure of poster's identity. *Newspaper Research Journal*, *38*(3). https://doi.org/10.1177/0739532917722977

Munn, L. (2019). Alt-right pipeline: Individual journeys to extremism online. *First Monday*. https://doi.org/10.5210/fm.v24i6.10108

Sample, I. (2020, January 13). What are deepfakes – and how can you spot them? *The Guardian*. https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

Schumpe, B. M., Bélanger, J. J., Moyano, M., & Nisa, C. F. (2020). The role of sensation seeking in political violence: An extension of the Significance Quest Theory. *Journal of Personality and Social Psychology*, *118*(4), 743–761. https://doi.org/10.1037/pspp0000223

Seth, S. (2023, September 11). The World's Top 10 News Media Companies. *Investopedia*. https://www.investopedia.com/stock-analysis/021815/worlds-top-ten-news-companies-nws-gci-trco-nyt.aspx

Somers, M. (2020, July 21). Deepfakes, explained. *MIT Sloan*. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Tandoc, E. C., Lim, Z. W., & Ling, R. (2017). Defining "Fake news": A Typology of Scholarly Definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Van Puffelen, M. (2021). Rechtsextremisme: Geweld met een rechtsextremistisch motie. In *DSP-groep*. DSP-groep. https://www.dsp-groep.nl/wp-content/uploads/18MP-Rechtsextremisme-DSP-2021.pdf

Van Wonderen, R. (2023). *Rechts-extremistische Radicalisering op Sociale Media Platformen*. Verwey-Jonker Instituut.

Van Wonderen, R. (2023). *Richtlijn / onderbouwing Radicalisering*. Verwey-Jonker Instituut.

Van Wonderen, R. & Peeters, M. (2021). *Werken aan weerbaarheid tegen desinformatie en eenzijdige meningsvorming. Evaluatie lesprogramma Under Pressure.* Utrecht: Verwey-Jonker Instituut. https://www.verwey-jonker.nl/wp-content/uploads/2022/04/120550_Werken-aan-weerbaarheid-tegen-desinformatie-eenzijdige-meningsvorming.pdf.

Wasike, B. (2022). When the influencer says jump! How influencer signaling affects engagement with COVID-19 misinformation. *Social Science & Medicine*, *315*, 115497. https://doi.org/10.1016/j.socscimed.2022.115497

Wolfowicz, M., Weisburd, D., & Hasisi, B. (2021). Examining the interactive effects of the filter bubble and the echo chamber on radicalization. *Journal of Experimental Criminology*, *19*(1), 119–141. https://doi.org/10.1007/s11292-021-09471-0