The Spread of Deepfakes in Digital Networks

Tommaso Tonello,[1] Aseniya Dimitrova,[2]
Livio Fenga,[3] Luca Biazzo,[4] Alessio Jacona[5]

[1] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
t.tonello@uu.nl
[2] Brand Media Bulgaria, Sofia, Bulgaria
a.dimitrova@economic.bg
[3] University of Exeter, Department of Management, Exeter, United Kingdom of Great Britain and Northern Ireland
l.fenga@exeter.ac.uk
[4] University of Brescia, Department of Economics and Management, Brescia, Italy
luca.biazzo@unibs.it
[5] Agenzia Nazionale Stampa Associata, Rome, Italy
alessio.jacona@gmail.com

This chapter explores the spread of deepfakes through social media platforms (e.g. X, Facebook and TikTok). By studying real-world case studies, such as political deepfakes or celebrity impersonations, the chapter illustrates how synthetic media exploit online engagement dynamics to reach massive audiences quickly. It then reviews current methods used to detect and track deepfakes, especially early-warning systems monitoring content spread patterns to flag potential deepfakes in real time, as well as novel research instruments developed as part of the SOLARIS project. The chapter then presents the role of traditional media in debunking and contextualising deepfakes, reflecting upon the challenges that AI-generated disinformation poses to journalists and media professionals. In this context, insights from SOLARIS' Use Case 2 are used to show how targeted interventions can slow the spread of harmful synthetic media. Finally, the chapter advocates for bottom-up AI education to frame digital citizens' needs and to foster their ability to engage with online synthetic content.

DOI
https://doi.org/
10.18690/um.feri.2.2026.2

ISBN
978-961-299-109-8

Keywords:
deepfakes,
social media,
disinformation,
statistical analysis,
journalism,
education

University of Maribor Press

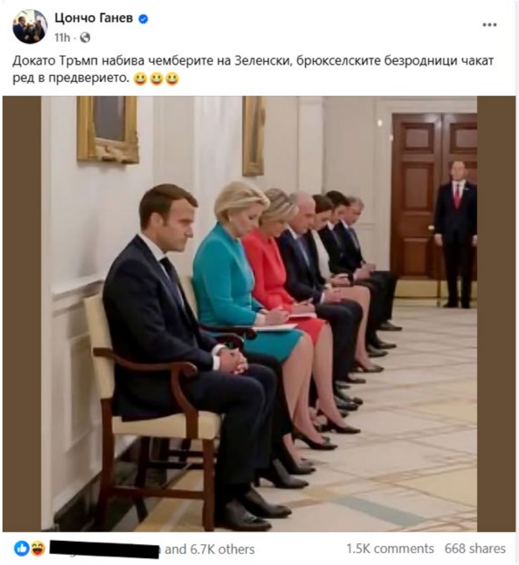
1 The Problem of Deepfakes and Social Media: How Deepfakes Go Viral

In an age where digital content moves at unprecedented speed, deepfakes have emerged as one of the most disruptive forms of synthetic media. Their increasing realism and accessibility raise pressing concerns about the manipulation of public opinion and democratic engagement, especially in politically sensitive contexts. This chapter investigates how deepfakes propagate across digital networks, with a particular focus on the architecture of platforms such as X (formerly Twitter) and Facebook. These environments, governed by engagement-driven algorithms and virality incentives, are especially susceptible to the rapid diffusion of deceptive content. Understanding these dynamics is essential to anticipating, detecting, and ultimately mitigating the societal risks posed by deepfakes. The analysed cases offer insights into how disinformation is packaged for viral spread. Ultimately, we point to the need for cross-disciplinary approaches, combining technical detection, network modelling, social media analysis, and media experts' insights, to map and counteract the spread of deepfakes and to disseminate relevant AI knowledge at the societal level.

This section details how deepfakes go viral on social media, drawing on examples from the U.S. and European political landscapes. The examples were picked based on their prominence and recency. It is beyond the scope of this chapter to analyse all cases and uses of disinformation using AI-generated content. Instead, we picked five examples that left a mark by reaching large audiences. Each of them is illustrative of the use of different social media channels based on the goals of misleading posts created or shared by online users. We then draw some conclusions on the mechanisms by which online network algorithms can enhance deepfake distribution.

1.1 The Case for the Obedient European Leaders

A most recent example comes from the context of the Russia-Ukraine war. It follows a meeting between European leaders at the White House on 18 August 2025, which took place as part of the peace-building efforts by the Trump administration. During the event, President Trump had lengthy discussions with Western leaders, including French President Emmanuel Macron and the European Commission President Ursula Von der Leyen, to agree on a common negotiating position.



Цончо Ганев
11h ·

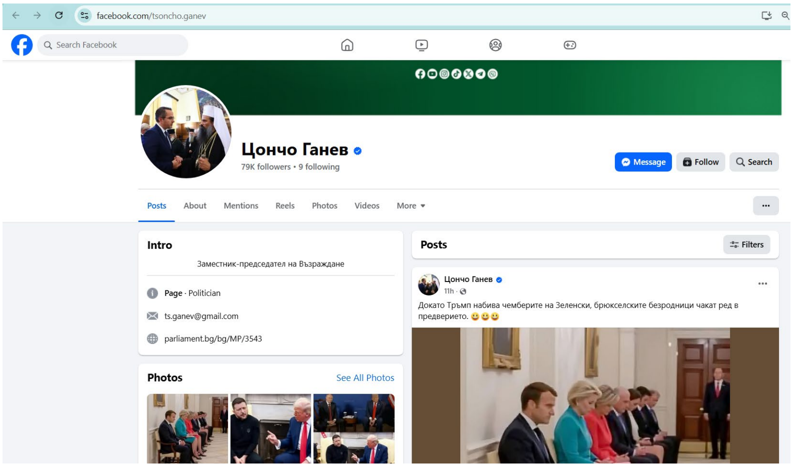
Докато Тръмп набива чемберите на Зеленски, брюкселските безродници чакат ред в предверието. 😃😃😃

[redacted] and 6.7K others 1.5K comments 668 shares

Figure 2.1: Detail of European leaders queuing to meet President Trump (deepfake)
Source: https://www.facebook.com/tsoncho.ganev.

However, a widely circulating deepfake image claimed to be taken on the day of the event portrays European leaders sitting obediently, heads down, waiting for the American President to return with instructions. The post claims to show the leaders waiting for President Trump to finish schooling President Zelensky. Whatever the interpretation, the message is clear: European politicians are portrayed as showing weakness, being sidelined by the great leaders of Russia and the USA, and they are only observers of important events in international politics. Only, this never happened, and there are clear signs that the image is fake.

The image has been circulating widely on Facebook and X, but the screenshot showcased above is taken from the Facebook page of a high-ranking pro-Russian politician from Bulgaria, Tsoncho Ganev, member of Parliament and vice-president of the pro-Russian Vazrazhdane (Revival) party, which maintains close ties to Putin's United Russia party, the two having recently signed a collaboration agreement. The post caption reads in slang: While Trump is schooling Zelensky, the barren Brusselers are waiting their turn in the lobby. Ganev is probably not the real author of the image, because the quality is low (suggesting he probably saw it somewhere and took a screenshot). Nevertheless, he started an information thread which became widespread in Bulgaria.



Цончо Ганев
79K followers • 9 following

Intro

Заместник-председател на Възраждане

Page · Politician
ts.ganev@gmail.com
parliament.bg/bg/MP/3543

Photos

Posts

Цончо Ганев
11h ·

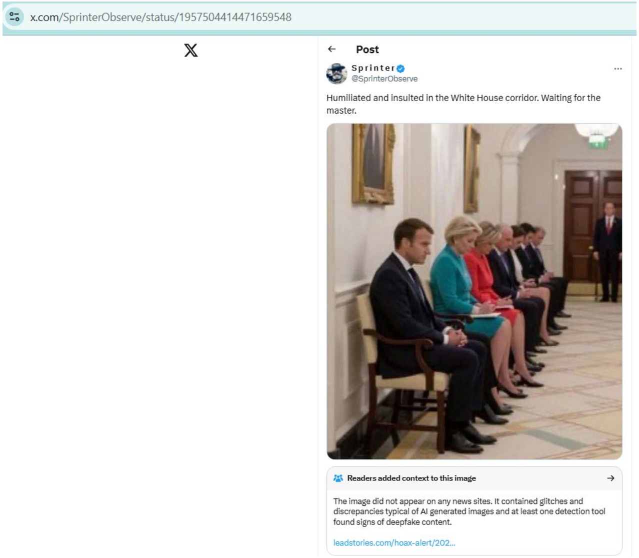
Докато Тръмп набива чемберите на Зеленски, брюкселските безродници чакат ред в предверието. 😃😃😃

Figure 2.2: European leaders queuing to meet President Trump (whole post on Facebook)
Source: https://www.facebook.com/tsoncho.ganev

Within hours, the image had amassed hundreds of shares and thousands of comments, mostly supportive, although it clearly is a deepfake – this can be seen by several inconsistencies, including a pair of legs with no body between the French and the EU Commission presidents, a mismatch between the outfits they actually wore that day and those shown in the image, a difference between President Macron's shoes, etc. At the time of writing, and despite several reports to Facebook that the image is false, it has not been taken down, nor has any context been added by the network to label it as a deepfake. The same politician has also shared the content on their X page, but since this network is not highly popular in Bulgaria, the effect it produced there was of a different magnitude.

A basic manual review of shares shows that among the profiles that have shared the image on Facebook, there are genuine profiles, largely pro-Russian supporters, official pages of political party structures, and many fake profiles (with fake images, low numbers of friends, mostly propaganda-style content). The post has also been shared in several Facebook groups publishing, among other things, anti-Western integration (for example, one that opposes Bulgaria's integration into the Eurozone), anti-establishment, and anti-George Soros content. This testifies to the importance of information bubbles on social media, safe spaces where we encounter mostly information that fits into our own worldviews and comes from sources that we consider safe and credible.

Meanwhile, on X, the same image shared by a verified user (called Sprinter Observe), with 770 k+ followers, has accumulated 104.8K views within a few hours and a similar number of shares. However, we can already see readers having generated a contextual note, saying that this is a fake image and explaining why.



x.com/SprinterObserve/status/1957504414471659548

Post

Sprinter
@SprinterObserve

Humiliated and insulted in the White House corridor. Waiting for the master.

Readers added context to this image

The image did not appear on any news sites. It contained glitches and discrepancies typical of AI generated images and at least one detection tool found signs of deepfake content.

leadstories.com/hoax-alert/202...

Figure 2.3: European leaders queuing to meet President Trump (X version of the post)
Source: https://www.facebook.com/tsoncho.ganev (Tsoncho Ganev on Facebook)

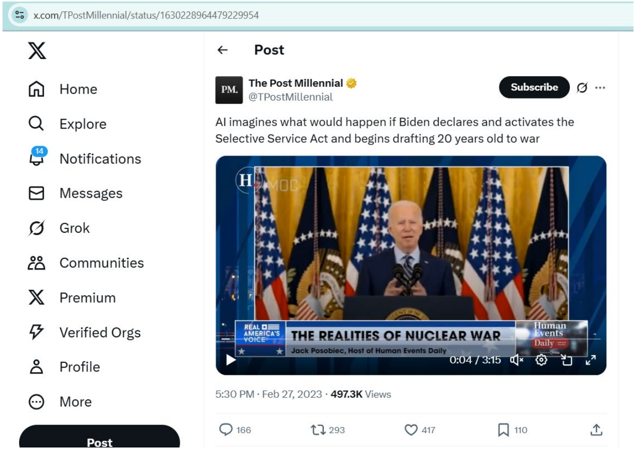
The post caption reads: “Humiliated and insulted in the White House corridor. Waiting for the master.” This indicates that the author of the post intended to present it as true. While in the first example, the author of the post is clear, a known political figure aiming to strengthen their fan base and solidify support behind pro-Russian views in a critical time, in the X case, there is not much information about the author of the post. It claims to be an independent media reporter, but there is no additional public data to associate it with someone’s identity. The only external link from the profile leads to a donation page. A reverse search of the profile image shows it is a portrait of Issam Zahreddine, one of the main commanders of Bashar al-Assad's army, killed in Syria in 2017, hence, not the real author of the post. This did not prevent the content from becoming viral, nor has it prompted the network to take down the profile or limit its exposure as being non-genuine.

This case might not be the most prominent example of social media use of deepfakes to harm, but it is pertinent and clearly shows the rapid spread of falsified content on Facebook, which can be re-shared with a lack of criticism and powered by influential figures from the political world and from the civic side itself. The fact that the posts have not been removed from their authors' profiles and no explanation for their authenticity has been given suggests that the intention has never been to inform, but rather to create a lasting impression. A large body of experimental literature shows that misinformation often continues to influence people even after it has been explicitly debunked - the so-called continued influence effect (Lewandowsky et al., 2012). Corrections reduce but frequently do not fully eliminate the influence of the false claim; in some circumstances, corrections can fail or even (rarely) backfire. Therefore, a falsified picture such as the example above would leave a lasting impression on the audience, and the longer it stays online, the stronger the impression. We simply cannot unsee a picture, even if we have later been made aware that it has been manipulated.

1.2 President Biden Calling for a National Draft to Defend Ukraine

Another interesting example, once again from the context of the war in Ukraine, comes in the form of a deepfake video circulated on X. It depicts then-President of the USA Joe Biden during a briefing calling for a national draft allowing for men and women from the States to be called to fight in Ukraine. One of the first appearances of this content occurs on X on 27 February 2023 by a news aggregator called The Post Millennial. The caption of the post clearly states that the video is AI-generated, only to depict a fictitious scenario. A commentator later in the video also confirms this is not a real event, but content that has been scripted and designed by the production. A more detailed check establishes that the new video was a doctored version of a video released by the White House on another occasion back in 2021.

The video is a relatively good deepfake, as it is somehow credible in the sense that it depicts something that many people feared might happen; the images and sound are also realistic, and only a deeper look into the gestures of Biden shows that something is wrong. The post has gathered a considerable number of views and shares, but it is nothing unusual, given that the page is a popular one with over 430k followers.



x.com/TPostMillennial/status/1630228964479229954

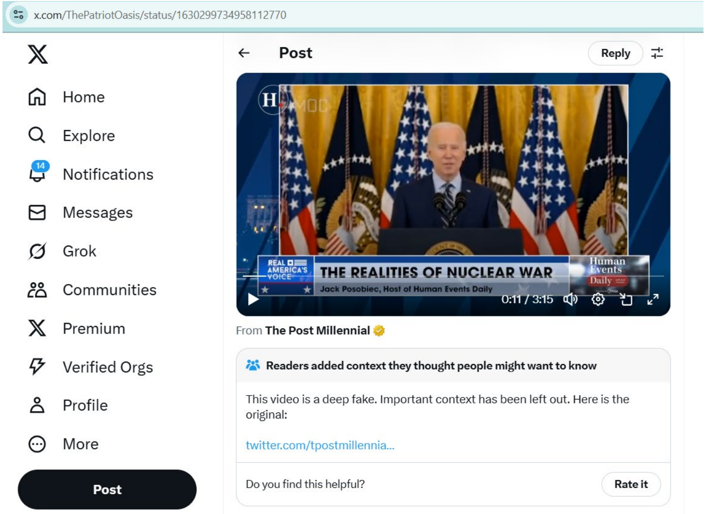
The Post Millennial
@TPostMillennial

AI imagines what would happen if Biden declares and activates the Selective Service Act and begins drafting 20 years old to war

5:30 PM · Feb 27, 2023 · 497.3K Views

Figure 2.4: Video of President Biden announcing a national draft
Source: https://x.com/ThePatriotOasis/status/1630299734958112770.

The situation becomes much more interesting, as the video has been re-shared (although with a very different caption) by another page on X, the Patriot Oasis. While it has a smaller fan base than the original, the post has now accumulated over 8 million views. The difference: it presents the video as if it were genuine, using words such as "BREAKING" for a stronger emotional effect. The fact that it comes from a patriotic page might also have contributed to this.



x.com/ThePatriotOasis/status/1630299734958112770

From The Post Millennial

Readers added context they thought people might want to know

This video is a deep fake. Important context has been left out. Here is the original:

twitter.com/tpostmillennia...

Do you find this helpful?

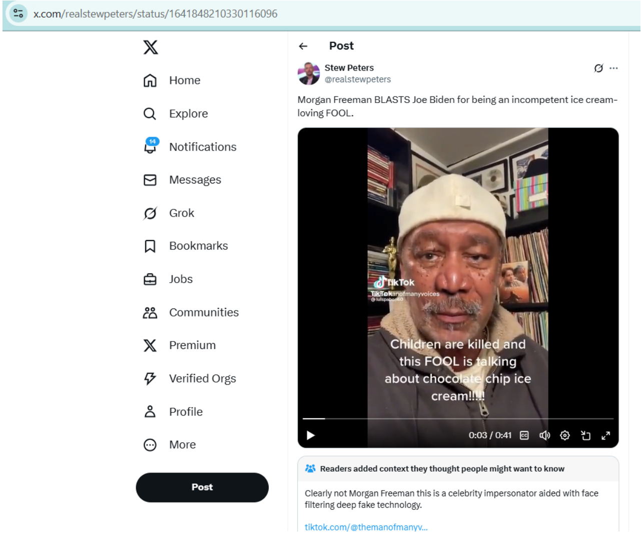
Figure 2.5: disclaimer flagging President Biden's national draft video as deepfake
Source: https://x.com/ThePatriotOasis/status/1630299734958112770.

This time, we can see that readers have added context explaining that the video is fake, but we do not know how many people noticed the warning and were influenced by it in the meantime. The different distributions of the same piece of content clearly show how disinformation spreads as fast as real news.

Growing scientific evidence shows that negative emotions, such as fear, anger, anxiety, and sadness, are systematically used on social media to amplify the spread of disinformation and, importantly, online engagement - to the benefit of social media platforms (Ali Adeeb & Mirhoseini, 2023).

1.3 Morgan Freeman calling President Biden a fool

Another example from the political sphere comes from the USA, but this time, there are two targets: the protagonist of the video, American actor Morgan Freeman, and Joe Biden, against whom the deepfake is addressed. The video depicts a poor-quality Freeman allegedly criticizing the President for being irrelevant in the situation of a mass shooting in the USA and calling for his removal from office. Originally, the video appeared on TikTok but was deleted: the post below comes from a repost of conservative radio host Stew Peters with the caption Morgan Freeman BLASTS Joe Biden for being an incompetent ice cream-loving FOOL.



Stew Peters
@realstewpeters

Morgan Freeman BLASTS Joe Biden for being an incompetent ice cream-loving FOOL.

Children are killed and this FOOL is talking about chocolate chip ice cream!!!!!

Readers added context they thought people might want to know

Clearly not Morgan Freeman this is a celebrity impersonator aided with face filtering deep fake technology.

tiktok.com/@themanofmanyv...

Figure 2.6: Morgan Freeman criticizes President Biden (deepfake)
Source: https://x.com/realstewpeters/status/1641848210330116096

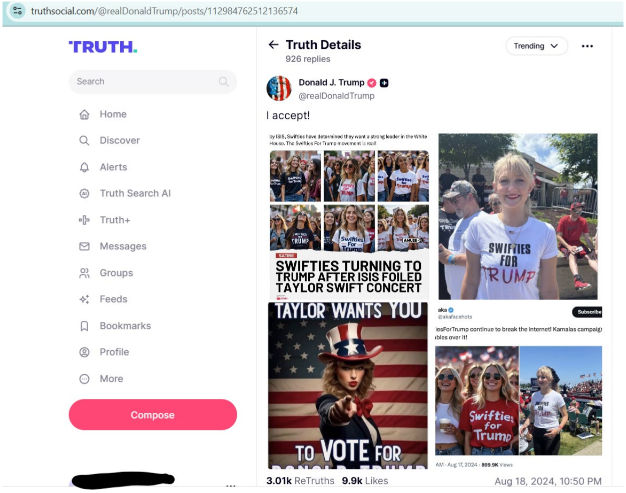
Once again, the author uses a well-known media technique to attract attention and evoke emotions: capital letters and dramatic words. The use of children in the text also evokes emotions, making use of a national tragedy to add another layer of criticism to the former President. This clearly has had an effect, as the post has gathered 5.3 million views and thousands of shares. Unsurprisingly, among the sharers, we find people expressing political partisanship, but also a lot of seemingly fake profiles. This time as well, however, many people also debunked the content. As for the poor video quality, the movements of Freeman appear very unnatural, as if a mask was superimposed on his face. What is more, if the video was genuine, one would expect it to be posted by its claimed author as well. However, the actor himself does not have a TikTok account.

This additional verification check is unlikely to be taken by most users, especially if they are emotional and are already prone to believing the suggested story about the President. In fact, as the idea of confirmation bias teaches us, people are more likely to accept the truth of news supporting their existing beliefs, while also discounting contradictory evidence. This becomes especially powerful on social media, where people often share headlines without reading them, relying on intuitive judgment rather than analytical thinking, especially when the content is emotionally charged or aligns with their views (Pennycook & Rand, 2019).

1.4 President Trump Endorsed by the Swifties

Another, more benign example can be observed on Truth Social – the social network of Donald Trump. It has been shared by Donald Trump himself in the context of his second electoral campaign. It is a compilation of screenshot posts from X users containing deepfake images of media articles and photos of young girls, seemingly fans of Taylor Swift, who demand a strong leader and are rallying against a Swifties for Trump movement. They also use capital letters in the caption and bait words such as “SHOCK”. The original posts have gained hundreds of thousands of views. Trump’s post is from August 2024, and it follows the cancellation of a Taylor Swift concert in Vienna due to possible terrorist attacks planned by ISIS. In reality, Taylor Swift had not endorsed Trump and had also criticized him publicly.

Trump has obviously combined a few posts and screenshots to steer public opinion in his favour, accompanying the post with a caption reading that he accepts being the strong leader in the White House. The post can contain some truth. Likely, the author of the original post is a Trump and a Swift fan, who has even crafted herself a t-shirt with the label Swifties for Trump. All other illustrative images, however, have obviously been generated with AI.



[redacted]

Truth Details
926 replies

Donald J. Trump
@realDonaldTrump

I accept!

by ISIS, Swifties have determined they want a strong leader in the White House. The Swifties For Trump movement is real!

SATIRE

SWIFTIES TURNING TO TRUMP AFTER ISIS FOILED TAYLOR SWIFT CONCERT

TAYLOR WANTS YOU TO VOTE FOR

aka
@akafacehots

iesForTrump continue to break the internet! Kamalas campaigr ıbles over it!

AM · Aug 17, 2024 · 899.9K Views

3.01k ReTruths 9.9k Likes

Aug 18, 2024, 10:50 PM

Figure 2.7: President Trump's post on Truth claiming endorsement from Taylor Swift's fans
Source: https://truthsocial.com/@realDonaldTrump/posts/112984762512136574.

The post has not accumulated that many views and shares, but it illustrated another possible use of deepfake content on social media: to fake support and endorsement of political candidates.

There is no evidence behind the intentions of Donald Trump, but there is solid scientific evidence showing that celebrity endorsements and influencer status can significantly increase the perceived credibility of fake news or misinformation, even when the content itself is misleading (Mena et al., 2020). A study using eye-tracking experiments demonstrated that articles featuring celebrity images and sensational headlines (fake news style) command more viewer attention than other content, even drawing attention away from the article's factual text. This signals a strong unconscious attraction to celebrity-linked fake content (Lazar & Pop, 2021). At the same time, it is known that celebrity amplification can cause real harm, something that has been documented multiple times during the Covid-19 pandemic, when influencers, including celebrities and wellness figures, played outsized roles in spreading anti-vaccination conspiracies, introducing personal narratives that increased engagement and made moderation more complex (Observatory, 2022).

1.5 Vladimir Putin talks to… Vladimir Putin

Finally, another example of a deepfake which has spread rapidly online, but this time for a different purpose: to educate the public, or rather, to convey a political narrative. The video comes from Russia and is state-sponsored. To address numerous rumours appearing in Western media, claiming that President Putin does not personally attend meetings, but uses doubles, the team behind the Russian president has decided to perform a media exercise and show how easy it is to be fooled by deepfakes. It shows real Vladimir Putin sitting in a studio with a live audience during his annual news conference. The president is looking at a screen, taking questions about policy from remote speakers. At some point, an AI-generated Putin lookalike appears, presenting himself as a student. He has the body and voice of Putin, and it therefore looks like the president is talking to his Doppelgänger. During the conversation, Putin's AI look-alike asks the president if he has a lot of doubles and his opinion about the dangers of deepfakes. The content originally appeared on national TV in December 2023 and only then spread to social media worldwide, making it impossible to track its exact spreading path. The numerous news headlines from large online media show that it made an impact. Alongside the purpose to inform and to spread fear that something happened to the Russian leader, this video also served the Russian-state propaganda goal of portraying Western media as biased against Russia, by using the very weapon Russia is usually blamed for using: disinformation.

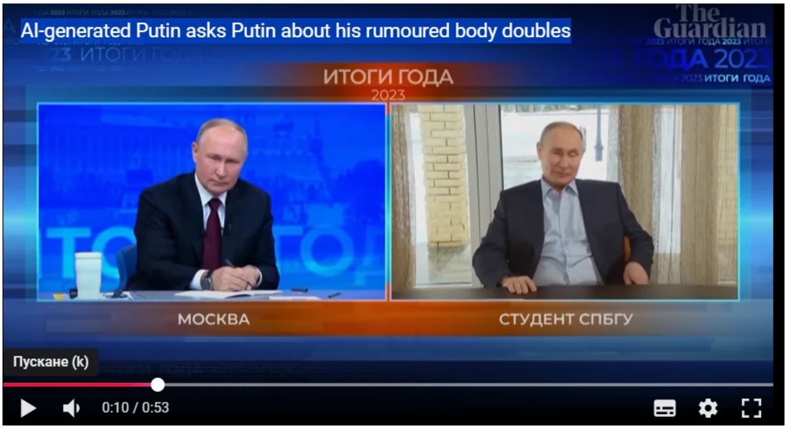


Figure 2.8: Putin talks to AI-generated Putin.
Source: The Kremlin via The Guardian
https://www.theguardian.com/technology/artificialintelligenceai/2023/dec/14/all.

1.6 Challenges

Most social media can become a vehicle of deepfake disinformation. Some of the key factors enabling this are the rapid content distribution, a trusted environment in closed groups, filter bubbles, echo chambers, anonymity, private chats, influencers, resulting in the empowerment of virtually all users to become media on their own.

Disinformation is an intentional act, with its authors usually choosing the best network depending on their needs. Engagement-driven algorithms of Facebook, for instance, keep showing us more of what we like, encouraging users to engage with similar content and causing stronger emotional reactions. Its large user base, which includes many users who are not used to detecting risk factors in digital environments, combined with current struggles to detect AI-generated disinformation and failure of automatic content moderation, makes Facebook an ideal ground for deepfakes disinformation. Platforms like X are doing better with flagging AI-generated disinformation and adding context, but the platform's dominant political and news orientation allows for politically motivated deepfakes to spread rapidly.

There are now many challenges to analysing how content spreads on social media to regular users or independent journalists. Previously easily accessible tools like CrowdTangle, a Facebook software allowing users to follow the spreading of online content, have been discontinued and replaced by less efficient and accessible alternatives (Gotfredsen & Dowling, 2024).[1] Notably, alternative tools for trend analysis and monitoring are available, but they are also expensive and usually require some degree of technical knowledge.

Most importantly, even if bots and fake profiles boost the distribution of a deepfake, a very concerning fact is that it is very often popular public figures, influential in the public space, who distribute deepfakes, exploiting emotions, patriotism, vulnerable groups, and sensitive social topics to serve their goals.

[1] Meta claims that Meta Content Library (MCL) is the new tool to provide high-quality data to researchers, while abiding by regulatory requirements for data sharing and transparency. However, reports claim that this tool is much less accessible, transparent and useful.

While social networks are incapable (or unwilling) to slow the spread of deepfakes, since their internal policies and one-size-fits-all interventions are proving too slow or inefficient, progress by experts promises to help tackle AI disinformation concerns. A step in this direction is represented by statistical approaches monitoring disinformation waves that, by identifying distinct, vulnerable populations, can then help to identify customized and more effective debunking interventions.

2 Statistical Approaches to Segmentation

The analysis of propagation dynamics and statistical detection models presented in this chapter provides the theoretical and technical framework necessary to interpret the case studies discussed in the previous section. While the latter examined the tangible effects of synthetic disinformation, such as the manipulation of public opinion through the falsified image of European leaders or the doctored video of President Biden, this section deconstructs the underlying mechanisms driving these phenomena. It becomes evident, for instance, that the virality of such content is not accidental, but rather the predictable result of the interplay between the engagement-driven algorithms described in the previous section and the emotional levers of fear or indignation that characterized those specific episodes.

Furthermore, the hybrid detection methodologies proposed in this section, grounded in sentiment analysis and time-series anomaly detection, directly address the critical vulnerabilities exposed in the previous examples. Where the human eye and traditional verification methods reached their limits against the visual hyper-realism of the Morgan Freeman deepfake or the rapid dissemination of falsehoods on Twitter, the statistical approach illustrated here offers a tool capable of identifying the latent traces of manipulation. Consequently, this section does not merely describe network operations: it proposes a methodological response to the systemic vulnerabilities exemplified by the narratives described previously.

As discussed in the previous section, the spread of synthetic media, especially deepfakes created with generative AI, has deeply changed the digital information landscape, creating serious challenges for truth, public debate, and democratic stability. What began as an innovative technology now enables the rapid and convincing spread of fake content, greatly strengthening disinformation efforts. Because of this, it is crucial to take a critical look at existing statistical methods, beginning with segmentation techniques that group people by their level of vulnerability, and moving toward advanced models that uncover the subtle social effects of AI-generated false information.

At the core of these developments is the need to view the digital information space as a complex ecosystem shaped by many different actors. These actors include individual users, each with distinct cognitive styles, emotional traits, and levels of trust in media, as well as collective agents such as social media platforms, algorithms, automated bots, and influential content creators. Together, they shape the speed, scale, and spread of synthetic media, including deepfakes and other forms of advanced misinformation.

This complexity creates the need for a comprehensive analytical framework integrating micro-level processes, such as individual susceptibilities, cognitive biases, and emotional responses, with macro-level systemic structures like networks and algorithmic affordances. Only by jointly examining these dimensions can researchers map how vulnerabilities emerge, disseminate, and embed in the digital milieu.

Advanced statistical modelling plays a key role in examining the diversity and variation within a population. The psychological foundations discussed in Chapter 4 will later explain how sociodemographic, motivational, and cognitive factors shape people's susceptibility to deepfakes techniques such as logistic regression, latent class analysis (LCA), factor analysis, clustering algorithms (used to group similar things together), and structural equation modelling (SEM) allow researchers to extract latent psychological and behavioural profiles from complex datasets. These tools identify distinct risk groups and reveal how interconnected beliefs, emotions, ideologies, and digital engagement cultivate susceptibility (Bhatnagar & Ghose, 2004; Kang et al., 2020; Outwater et al., 2003; Verma, 2013; Yan et al., 2018).[2]

However, current models have some limits. They often look at only a few factors and rely too much on data from Western countries. To make them more useful, researchers need to include data from more regions and cultures and use long-term, cross-platform studies to track how people's vulnerability changes over time.

[2] Logistic regression statistical method that predicts the probability of something happening and turns that into a yes/no decision. LCA is used to find hidden groups (or "classes") within a set of people (or items) based on their answers, behaviours, or characteristics. Factor analysis is used to find underlying patterns or "factors" in a large set of variables. It helps researchers understand which variables go together and what hidden dimensions explain them. Finally, SEM is a powerful statistical method used test complex cause-and-effect relationships between observed and hidden (latent) variables, all at once, in a single, comprehensive model.

Therefore, building an effective and lasting response to AI-driven disinformation requires collaboration across different fields, combining insights from psychology, statistics, computer science, and socio-political studies. This well-rounded approach is crucial for identifying where people are most vulnerable and developing evidence-based strategies that strengthen democratic resilience in a constantly changing information environment.

2.1 Statistical Modelling Approaches for Studying the Impact of GenAI Content and Fake News

Recent advances in statistical modelling have substantially deepened our understanding of the multifaceted and often subtle ways in which AI-driven synthetic misinformation spreads, affects, and reshapes different societal groups. Researchers now use a wide range of sophisticated quantitative methods to uncover the multiple, context-dependent factors that drive susceptibility, moving beyond basic descriptive analyses toward detailed modelling of influence networks, belief formation, and behavioural dynamics (Sæbø et al., 2020).

Together, these statistical methodologies unlock unprecedented insights into the complex factors driving the spread and societal impact of AI-generated fake news. They allow researchers to map intricate networks of influence, which are often shaped by automated bots, coordinated influencer campaigns, and opaque platform algorithms, and translate this knowledge into practical, evidence-based solutions. These solutions range from carefully targeted media literacy programs designed for specific risk groups to predictive tools that identify emerging vulnerability clusters, to real-time content detection and moderation systems that can interrupt misinformation cascades at critical points, as well as adaptive regulatory measures that help platforms and policymakers respond quickly and effectively to the evolving disinformation landscape.

The true power of statistical tools lies in their ability to integrate theory and practice: turning conceptual understanding into evidence-based, context-sensitive interventions that help civil society and institutional actors detect, anticipate, and counter the harms caused by synthetic media. In an era defined by the rapid evolution of generative AI and the growing sophistication of synthetic content, only a continuously adaptive, data-driven, and theoretically grounded approach can protect the integrity of knowledge and strengthen democratic resilience in digital public spaces, thereby safeguarding the foundations of informed citizenship in the twenty-first century.

At the forefront of this endeavour stands logistic regression, a versatile statistical tool pivotal in isolating and quantifying individual-level risk factors (Shete et al., 2021). Variables such as age, educational background, ideological leanings, and media consumption patterns are no longer treated as mere demographic markers but are examined as dynamic mediators and moderators situated within complex psychosocial ecosystems. For instance, the protective influence of education may depend heavily on a person's digital literacy, while political ideology can influence news consumption and openness to misinformation in complex, non-linear ways. By incorporating these factors within interacting cognitive and sociocultural networks, logistic regression provides a nuanced understanding of how vulnerabilities emerge, showing how individual predispositions interact with structural exposures to increase susceptibility to fake news.

Latent class analysis (LCA) expands analytical possibilities by moving beyond predefined groups to reveal hidden subpopulations whose vulnerabilities stem from unique combinations of beliefs, emotional traits, and media engagement patterns (Shen & Wu, 2024).

This method is particularly effective at revealing the fluid and overlapping nature of audience segments that cannot be easily captured by simple demographic or psychographic categories. For example, LCA can identify clusters of users whose exposure to synthetic media is shaped by the combined effects of cultural norms, peer influence, and algorithmically curated content, together creating hidden vulnerability profiles. This approach reframes susceptibility not as a fixed individual trait but as a dynamic interaction of self-concept, social identity, technological mediation, and the broader networked environment, highlighting the need for innovative segmentation models and precisely targeted interventions.

Adding another layer of methodological sophistication, structural equation modelling (SEM) allows researchers to estimate both direct and indirect causal pathways connecting a complex set of cognitive, emotional, and socio-structural variables (Tahat et al., 2022). SEM is particularly effective at analysing the recursive and often bidirectional feedback loops found in digital misinformation ecosystems. It maps complex relationships, such as how media trust directly influences credulity, or how ideological alignment affects the emotional impact of deceptive content. For example, SEM can model how initial acceptance of a deepfake sparks emotional arousal, which then increases selective sharing and fosters attitudinal polarization within networked communities. This level of analytical detail is essential for understanding the self-reinforcing dynamics that drive the spread and lasting impact of synthetic media among digitally connected audiences.

2.2 Case Study: Early Detection of Fake News through a Hybrid Statistical Framework

Within the SOLARIS project, we developed an innovative hybrid statistical model designed to enhance the identification of AI-generated fake news. This approach integrates diverse analytical techniques to improve both the accuracy and timeliness of detecting synthetic misinformation within dynamic digital environments.

Our methodology operates on two complementary levels. The first one focuses on analysing the emotional tone of news articles using sentiment analysis (Mohammad & Turney, 2013). Here, we measure the expression of key emotions such as fear, anger, sadness, and trust throughout a text. It is consistently observed that fabricated news exploits emotional manipulation, often intensifying negative emotions like fear and anger to capture reader attention and influence perceptions. By assessing patterns of emotional intensity and variability, we distinguish characteristic differences between real and fake news; as suggested in the previous section, authentic journalism generally maintains a balanced and steady emotional tone, whereas misinformation reveals abrupt spikes in distressing sentiments.

The second level concentrates on behavioural data, specifically analysing public engagement through online search trends. For instance, we monitored monthly search interest for the term “nuclear” spanning from 2004 to 2025 (see Figure 2.9 below). Sudden, anomalous surges in search volumes signal potential misinformation events or coordinated disinformation campaigns igniting public concern.

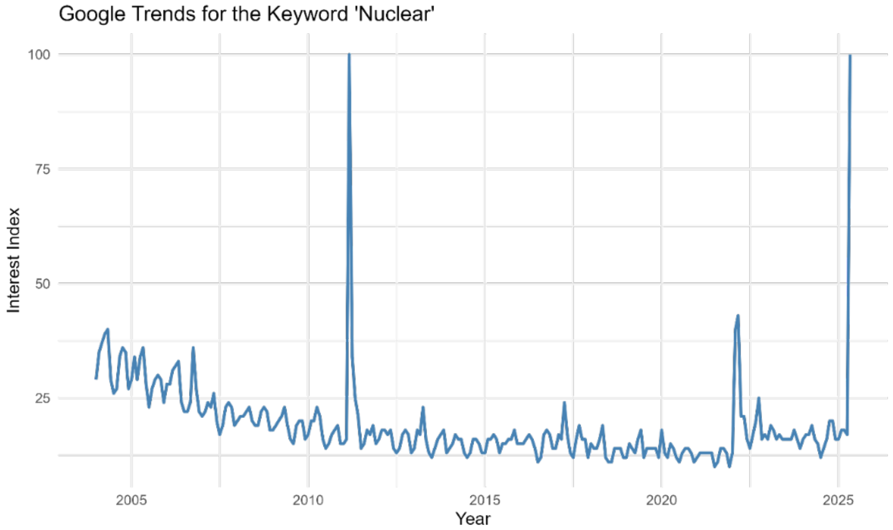


Figure 2.9: Monthly Google Trends data for the keyword nuclear (2004–2025). The final observation is artificially adjusted to simulate an anomalous spike.
Source: Fenga and Biazzo, 2025.

To robustly detect such anomalies, we deploy multiple forecasting models, including traditional time series techniques, such as Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ETS), alongside advanced machine learning models like the Extreme Learning Machine (ELM) neural network (Chatfield et al., 2001; Shumway & Stoffer, 2017; Wang et al., 2022).[3] We further enhance reliability using bootstrap resampling methods to generate confidence intervals, defining expected "safe zones" of variation against which real-time observations are evaluated (Hesterberg, 2011).[4] Once observed search frequencies exceed these bounds, the system flags a possible fake news event.

In experimental evaluations, we constructed a dataset comprising 20 genuine news articles alongside 5 AI-generated fake news pieces, paired with corresponding Google Trends data. Artificially injecting anomalous spikes into the search data, we tested the system's detection efficacy. The sentiment analysis reliably separated

[3] ARIMA is used to predict future values in a time series – like stock prices, weather, or website traffic – based on past data. ATS is a method for forecasting future values in a time series by giving more weight to recent observations and less weight to older ones. Finally, ELM is a type of artificial neural network used for classification or regression tasks – basically, for predicting outcomes or categorizing data

[4] Bootstrap resampling allows researchers to estimate the reliability of a statistic by repeatedly sampling from data, even if they do not know the underlying population.

fabricated from authentic content, evidencing higher levels of negative emotion and volatility in fake news. Concurrently, all forecasting models successfully and synchronously detected the synthetic anomaly, without false alarms during baseline periods, confirming the system's sensitivity and robustness.

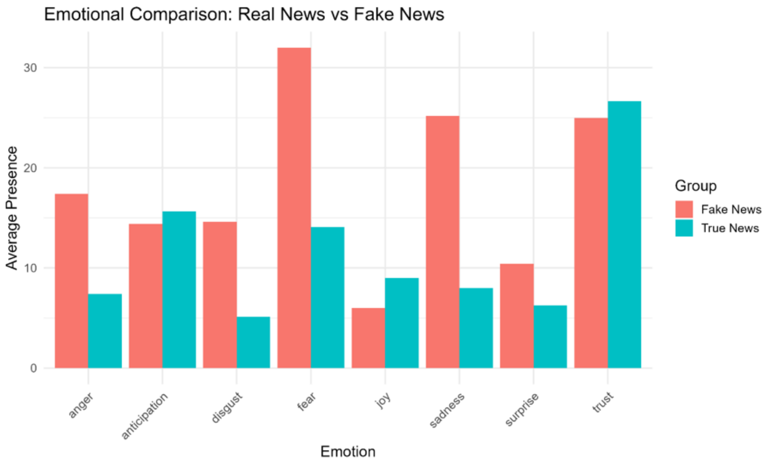


Figure 2.10: Relative emotion activation frequencies. Fake news intensifies fear, anger, and sadness.
Source: Fenga and Biazzo, 2025.

This dual-layered framework offers a potent early-warning tool against the proliferation of fake news. By uniting semantic emotional insights with behavioural metrics derived from real-time search activity, the model facilitates timely alerts for journalists, fact-checkers, and digital moderators, allowing for swift responses to emerging disinformation. Importantly, it is conceived as an augmentation rather than a replacement of human expertise, providing prioritized signals that guide investigative and corrective action. Its modular design permits adaptation across diverse languages and topical domains, enhancing its versatility and broad applicability.

2.3 Future Directions

Looking forward, combining advances in theory, statistics, and computation creates a strong research agenda to address AI-driven synthetic misinformation. As generative technologies increasingly blur the line between reality and fabrication, current models show important limitations and highlight the need for interdisciplinary approaches. Understanding vulnerabilities will require integrating psychological, behavioural, technological, and socio-political factors, as well as conducting long-term and cross-cultural studies. Real-time analytics and advanced natural language processing can support predictive and responsive interventions, helping policymakers and platforms act quickly when misinformation threatens social cohesion and democracy. At the same time, robust ethical frameworks and regulations are essential to protect privacy, rights, and public trust amid widespread digital manipulation. By building an adaptable, integrated framework that combines diverse data sources and methods, we can strengthen societal resilience against fake news and safeguard the integrity of public discourse and democratic institutions in this fast-changing digital era.

3 Detecting deepfakes on social media: the perspective of journalists and press agencies

For journalists and especially for freelancers, who often work alone with limited resources and under tight deadlines, the rise of deepfakes represents one of the most daunting and complex challenges faced in recent years; the same years in which an unprecedented technological revolution has profoundly transformed the world of information and, with it, the way the public reads and understands the present (Sohrawardi et al., 2020).

First came the pervasive spread of social networks such as X (formerly Twitter), Facebook, or Reddit: platforms whose algorithms decide what we see and when, based on criteria that are anything but transparent. These platforms have radically changed the way news is consumed, polarising opinions and systematically promoted "viral" content that generates engagement and, with it, valuable data for the very companies that produce and monetize these social networks. At the same time, the success of instant messaging systems such as WhatsApp, Telegram, or (to a lesser extent) Viber and Signal has created new spaces for exchange and sharing, such as channels and groups, where all kinds of content, including deepfakes, can be shared and reshared virtually without control (Al-Khazraji et al., 2023).

Now, adding to this landscape already extremely complex for journalists to decode, comes the unstoppable and rapid evolution of AI tools capable of generating fake audio and, above all, video content that is increasingly realistic, carefully crafted to go viral. It is a perfect storm putting great strain on a profession built on testimony and fact-checking.

The risk for journalists, and especially for freelancers, is twofold: on the one hand, there is the danger of falling into the trap after receiving an apparently authentic and relevant video, audio clip, or image (such as a fragment of a private conversation between politicians or an inconvenient admission by a public figure) and relaying it, thus becoming an unwitting cog in the disinformation machine. The urgent need to "stay on the story" and be the first to publish represents a shared necessity for both freelance and editorial journalists, with the major difference being the absence of a structured editorial team for cross-checking information for the former. A difference that can play a decisive role in the fight against disinformation and hinder professional integrity. The result: reputational damage that, for an individual professional, can be irreparable.

On the other hand, there is a subtler but equally insidious challenge: hyper-scepticism. When everything can be fake, verification work turns into an exhausting investigation. While the pillars of journalism, such as cross-checking authoritative sources or analysing context, remain the foundation of reporting, when every audio or video file becomes suspect, verification requires a process that drastically slows down the workflow, all while the "news" spreads uncontrollably across social networks. It is no longer just about cross-checking sources or verifying a witness's credibility, and about analysing a file's metadata and hunting for micro-imperfections in a video, such as an unnatural blink, a strange blur along the edge of a face, or inconsistent lighting. These details are becoming increasingly difficult to spot due to the progress of generative AI, as shown for instance, by the recent release of Veo 3, Google's video generator based on Gemini AI, which has "broken the silence barrier" by adding audio to ever-higher-quality images.

Fortunately, the same AI that creates the problem also provides part of the solution. Today's freelance journalist must necessarily combine a nose for news with technological competence. There exist AI tools specifically designed to detect deepfakes, and information professionals must learn to use them just as they once did with a notebook. Platforms such as Reality Defender, free software such as Deepfake-O-Meter, IdentifAI, or Sentinel (more suitable for companies and institutions), for example, analyse multimedia files submitted to them in search of digital artifacts and inconsistencies invisible to the human eye (Stephen, 2025).

Others focus on discrepancies between mouth movements (visemes) and spoken sounds (phonemes), a detail almost impossible to counterfeit perfectly.

However, the possibility of escalating (allegedly) fake news to other members of the editorial team points to the fact that, surprisingly enough, technology represents a last resort. These tools represent valuable support, but they cannot replace (and probably never will) human judgment and established journalistic practices: editorial journalists themselves tend to first cross-check with other sources reporting on the news, leveraging their newspaper's connections. By leveraging contacts with other newspapers, press offices, spokespersons, institutional social media profiles, and so on, editorial journalists are able to determine whether events depicted through a deepfake actually took place in the real world.

Second, journalists look at the context of the news. For instance, in case a public figure (such as a politician) were to give a speech that does not resonate with their known stance on the topic, say, a climate change denial message from activist Greta Thunberg, journalists may already flag the news as suspicious and, once again, check with other sources.

Finally, technical features of the video may be highlighted as suspicious by the expert eye of journalists, who may, for instance, detect discrepancies between mouth movements and spoken sounds, details almost impossible to convincingly counterfeit. Only at this point may editorial journalists resort to the help of detection software to analyse media content and determine whether the video depicts real or made-up events. This is the case, for instance, when journalists cover war areas, where it is difficult to cross-check with other sources or to extrapolate enough information from the context where events unfold.

In contrast to editorial journalists and the resources available to them, freelance journalists are able to resort to a multi-level approach: technology for initial screening, followed by a critical contextual analysis that only a journalist with the right expertise can provide. The fundamental question for both editorial and freelance journalists, however, remains the same: *cui prodest?* Who benefits from the spread of that false content? Then, as always, the process continues by cross-checking the news with known facts, testimonies, and primary sources. In short, navigating this constantly evolving landscape requires a new form of "augmented journalism": freelancers (as well as newsroom journalists) must become more meticulous, more transparent in their verification process, and above all, humbler. They must be ready to admit what cannot be verified with certainty and to explain to their audience the complexities of an information ecosystem where distinguishing between truth and falsehood has become the new, crucial challenge to overcome.

3.1 Use Case 2 – The SOLARIS Project Disinformation Event

Pursuing the goal of empowering journalists with relevant tools and skills to combat AI disinformation, the SOLARIS consortium organized a brainstorming session at ANSA's headquarters in Rome, involving journalists, communication experts, institutional representatives, researchers, private companies" professionals, and different stakeholders from the information sector. More specifically, the objectives of the event were as follows:

- collect feedback on how ANSA journalists detect and manage deepfakes in their daily work,
- co-design mitigation strategies, and
- formulate concrete recommendations to address "infodemics."

During the two days in which the roundtable debate took place, participants attended an editorial meeting to closely observe the daily working process of ANSA journalists of the newspaper agency's key activities. This allowed to witness the established processes and criteria by which ANSA decides which stories to cover and how to develop their reporting. Following the editorial meeting, a group of senior ANSA journalists was shown three deepfakes created specifically for the event: the goal was to assess their reactions and response procedures, as well as to identify possible gaps in current practices.

The debate then expanded into a session involving experts and the different kinds of stakeholders mentioned above, who started by identifying different types of AI-generated disinformation and their varying implications. Subsequently, the working group turned to the search for solutions, reflecting on the role of human beings in using their professional experience to combat disinformation and on the possibility of fighting fire with fire – that is, using AI to detect fake news, to promote digital literacy, and to create counter-narratives against deepfakes disinformation.

The event concluded with an interactive session in which ANSA journalists further discussed with experts the potential of detection tools and the adequacy of current laws and regulations targeting online disinformation.

3.2 Traditional Journalism vs. Deepfakes

The good news, then, is that professional journalism (especially with the support of the resources and practices of editorial settings), with its layered processes and models, already has many effective tools to counter deepfakes. The SOLARIS roundtable, in fact, highlighted a multi-level verification approach to identify and neutralize any false or manipulated content, including deepfakes. This process does not rely on a single tool, but rather on a combination of technical analysis, in-depth contextual knowledge, and rigorous journalistic principles.

The initial analysis of suspicious content often starts with superficial warning signs, such as evident imperfections in terms of context (missing or incorrect source logos), content (such as, for instance, a politician expressing a political stance incoherent with their long-held political beliefs), or obvious technical errors, like poor synchronization between audio and video. However, participants present at the brainstorming session stressed that the technical quality of a video is neither the only nor the most important evaluation factor: eventually, the true core of their defence strategy is keeping the human component at the forefront of technology use to tackle disinformation: journalistic experience makes the difference. Deep knowledge of specific contexts, sources, and public figures generally enables journalists to detect anomalies that an algorithm or an inexperienced eye would not be able to catch.

The network of regional correspondents and collaborations with other international news agencies (such as the BBC) acts as a cross-checking mechanism, essential for validating doubtful information, although editorial journalists argued they would not have cross-checked with other critical sources to verify the news, since technical, content, and contextual details all strongly pointed to the made-up nature of the videos analysed. Ultimately, ANSA journalists argued that the strongest defence lies in the core principles of journalistic work. Editors reiterated that source attribution is a fundamental and non-negotiable requirement. In an era of viral disinformation, the newsroom deliberately chooses to prioritise accuracy over speed, a principle that translates into the need to verify every story through direct contact with sources and to always seek multiple confirmations before publication.

Emerging from SOLARIS discussions, the key steps journalists may take against deepfakes can be summarized in the following order:

- The ability to cross-check online information with other media outlets or relevant institutions is at the heart of debunking disinformation.
- The content of deepfakes may provide very important hints: if the content is plausible, journalists need to leverage on their expertise to verify whether there exist inconsistencies in the message conveyed through the video.
- The context in which a video is set also delivers key insights about the content's credibility. With context, technical details (and journalists' ability to recognize them) become critical to detect fake news. Additionally, war contexts make videos more difficult to cross-check.
- Finally, supporting experts to identify technical inconsistencies, detection technologies may complement traditional processes with modern verification tools, including detection software based on AI.

The SOLARIS event also underlined the importance of a clearer taxonomy of disinformation. The discussions highlighted the crucial importance of distinguishing between "disinformation" and "AI-generated disinformation" – the latter encompassing video, audio, or written sources at an output-intensive pace compared to traditional disinformation – and "misinformation," the unintentional sharing of what is believed to be true, as well as "malinformation," which amplifies disinformation with defamatory intent. From the debate it emerged the need to differentiate "harmful content" according to its degree of risk.

Finally, among the critical issues that emerged from the dialogue between journalists and experts was also a worrying decline in public trust towards traditional media. To rebuild this trust – the panel suggested – it is essential to actively involve citizens rather than imposing knowledge from above. This can also be achieved by focusing on coaching professionals and end-users to understand the positive impact of generative AI on disinformation, which aims to use AI to detect deepfakes and generate content to develop counter-narratives to false news. More broadly, media literacy campaigns were recognized as a crucial tool to restore public trust and prepare citizens to navigate an increasingly complex information landscape.

4 Mitigating: Slowing the Spread

In the recent generative AI (genAI) wake, social scientists have pointed to the skill-replacing threat of AI technology over its skill-enhancing potential: people's ability to develop essential skills such as critical reading and structured thinking is hindered by the possibility to delegate tasks to AI tools, which makes education-related efforts appear redundant. Among other things, this translates to individuals being ill-equipped with the necessary knowledge to identify and react to online disinformation (Arribas et al., 2025). The affirmation of deepfakes as increasingly trustworthy visual content magnifies disinformation risks related to human-artifact interaction in the online context.

Citizens' inability to learn about and defend themselves from deepfakes hinders their status as rights-holders, eroding their capacity to self-advocate for the principles of transparency, privacy, and accountability. At the same time, deepfakes risk weakening democratic participation, widening social gaps by increasing the digital divide (Lythreatis et al., 2022). Against the backdrop of AI as a vector of technological disruption, experts have stressed the importance of democratising the values behind the introduction of AI tools: if citizens are to benefit from social media platforms and AI tools as a means for enhancing democratic engagement in the online context by combating disinformation, better inclusion of most diverse categories of citizens is most desirable in order to help identify socially critical AI problems (Corrêa & Oliveira, 2021).

However, the bottom-up approach must also be matched by efforts at empowering citizens with relevant knowledge on AI and deepfakes. By stressing the peculiarities of AI as a fast-changing technology, the limits of top-down regulatory approaches and institutional initiatives, the role of AI education as a precondition for enhancing the fight against AI-generated disinformation and strengthening individual rights in the online context is advocated for.

Economides (1996) and Birke (2009), focusing on Information and Communication Technologies, show that as more people adopt a network technology, its performance improves (Birke, 2009; Economides, 1996). AI systems exhibit this network externality too: the larger the data network they access, the more intense their training (LeCun et al., 2015; Panno et al., 2023). Learning-oriented algorithms nonetheless tend to go beyond what network technologies traditionally envisage in terms of spillover effects: in this case, the network features dramatically increase AI's ability to autonomously enhance its output (Levine & Jain, 2023). This, of course, also improves deepfakes' ability to mislead. The possibility to quickly create increasingly trustworthy deepfakes interacts with the global reach of world-famous platforms, such as those owned by Meta, which have occasionally contributed to political misinformation and disinformation dynamics (Acemoglu et al., 2025).

These problems have been approached by tightening the regulatory stance of national institutions. The EU context is usually taken as a benchmark comparison, considering the proactive regulatory stance the 27 have taken to address these problems. Legislative projects such as the AI Act and the Digital Services Act (DSA) have focused on preventing the introduction of AI technology deemed dangerous for end-users and on extending accountability of online platforms in terms of illegal and harmful content that may circulate through their digital environments. These initiatives mostly focus on engaging with technology producers, setting normative standards for the production of safe AI services. Alongside binding documents, the EU has also attempted to encourage voluntary compliance to safe information standards through the 2022 Strengthened Code of Conduct on Disinformation – integrated in the DSA in 2025 (European Commission, 2025). Such legal documents, however, do not yet appropriately tackle laypeople's AI education and critical skill development. Communication experts and journalists are therefore left to bridge the AI-generated information gap by either flagging fake content or by fact-checking the content of deepfakes (Painter, 2023). Forja-Pena et al. (2024) nonetheless stressed how newspapers are currently navigating the challenges posed to their working category from AI, investigating the ethical and efficient use of AI technologies to contrast disinformation and to help produce quality information (Forja-Pena et al., 2024). At the same time, they also highlight the lack of adequate technological literacy to tackle online disinformation and assist journalists in their jobs of quality reporting. Nonetheless, they also highlight the lack of adequate technological literacy to tackle online disinformation and assist journalists in their jobs of quality reporting. This represents a notable shortcoming in the fight against online misinformation, disinformation, and malinformation.

Even though AI education represents an urgent goal to be pursued in the context of combating disinformation, the delay in dissemination programmes stems from the ongoing debate on what constitutes relevant AI knowledge (Hermann, 2022; Kandlhofer & Steinbauer, 2018; Long & Magerko, 2020; Mikalef & Gupta, 2021):

what are the necessary notions to navigate a rapidly changing, self-enhancing technology? Given the dynamic nature of AI, would a theoretical and general preparation represent a better option than practical, AI tool-specific knowledge?

In the attempt to identify helpful AI notions, there exist governmental initiatives that have promised to prepare civil society to engage with AI tools and to promote political participation and the upholding of democratic values for digital citizens. By collecting citizens' input, such initiatives aim to inform the government's ability to support and provide adequate education and solve context-dependent problems of GenAI applications. A relevant instance of this political experiment comes from the Kingdom of the Netherlands, where the "Government-wide vision on generative AI of the Netherlands" advocates for country-wide resilience to AI-related challenges (Zaken, 2024). The resort to civil debate initiatives, such as the AI Parade, aims to collect data from citizens' experiences with AI technology, to articulate the goals of an AI education whose necessary knowledge is framed directly by digital citizens' needs.

Although the Dutch initiative does not revolve around the specific topic of AI disinformation, the constructivist approach of societal dialogue represents an important attempt at closing the information gap, at pursuing timely AI education, and at safeguarding democratic functions and norms. Providing citizens with the opportunity to share hands-on AI knowledge and to voice the expectations with respect to the introduction of different kinds of AI products and services is an unavoidable step, and it has been recognized as such by international stakeholders, even if this dialogue has mainly been understood from the perspective of preventing a worsening of the working conditions in relation to the introduction of AI (Cazes, 2023; Krämer & Cazes, 2022). Still, better regulation from institutions and enhanced cooperation by social media platforms are understood as the necessary and sufficient condition, or to the very least as the most urgent measure, to protect digital citizens and democratic institutions, with no complementary role envisaged for societal dialogue, AI education, and knowledge-sharing on online experiences (Painter, 2023; Pawelec, 2022).

Nonetheless, AI knowledge sharing is pivotal to the debate on a human-centred AI – that is, an ethical introduction of AI tools that enhance human capabilities rather than substituting them – and to the current regulatory focus behind strengthening democratic values and fostering ethical technological innovation (Khutsishvili, 2024). Therefore, the pursuit of civil debate and of knowledge sharing represents not a complementary and necessary component of tackling AI-generated disinformation, but an intrinsic element to the regulatory efforts and the scientific debate surrounding GenAI. Promoting a bottom-up AI education allows to tackle the legislative gap, to enhance efforts by journalists and fact-checking institutions, and to empower digital citizens to defend their rights.

5 Concluding Remarks

Deepfakes spread rapidly on social media by exploiting emotional responses, platform algorithms, and the authority of influential figures. The case studies examined illustrate how synthetic media can distort political discourse, cultural narratives, and public trust, often leaving lasting impressions even after exposure is corrected.

The statistical models and hybrid detection frameworks developed under SOLARIS represent innovative dual-layer research tools that merge computational linguistics with predictive analytics to detect disinformation patterns in real time. Specifically, our framework integrates sentiment analysis algorithms, which map emotional signals in text and identify manipulative spikes in fear, anger, or distrust, with advanced statistical forecasting models such as ARIMA, Exponential Smoothing (ETS), and machine learning techniques that track abnormal patterns in public engagement data. For example, if reports of an alleged "nuclear incident" emerged, the system would simultaneously analyse the emotional tone of the content against established thresholds while monitoring surges in Google search activity that exceed statistical confidence limits. These combined signals generate quantitative alerts, allowing experts to prioritise potentially fabricated content before it spreads widely. In doing so, this approach shifts disinformation detection from reactive fact-checking to proactive monitoring, functioning as a comprehensive "statistical radar" that unites textual manipulation analysis with audience behaviour across multiple languages and topics.

While statistical models and hybrid detection frameworks offer promising tools for identifying vulnerabilities and anomalous patterns, they remain limited by technological, cultural, and methodological constraints. Journalists, particularly freelancers, face a dual challenge: avoiding uncritical amplification of deepfakes while also resisting hyper-scepticism that undermines timely reporting. Evidence from SOLARIS activities underscores the enduring importance of human expertise, contextual knowledge, and professional standards as safeguards against manipulation. Effective mitigation requires an integrated strategy combining advanced detection tools, enhanced media literacy, regulatory frameworks, and stronger accountability mechanisms for platforms. Persistent obstacles such as filter bubbles, opaque algorithms, and declining trust in traditional journalism complicate these efforts. To tackle such challenges and safeguard democratic discourse in the digital age, empowering citizens to critically engage with digital content involves yet another key stakeholder in the fight against disinformation.

End notes

Aseniya Dimitrova led the overall structure, writing, and editing of the chapter, and prepared the Introduction, the How Deepfakes Go Viral, and the Conclusion sections. Tommaso Tonello provided detailed feedback and contributed to editing the chapter. He authored the section on Mitigating the Spread of Deepfakes on Social Media, co-authored other parts of the text, and supervised reference formatting for the entire chapter. Luca Biazzo and Livio Fenga jointly conceptualized the statistical framework and authored the sections on Statistical Approaches to Segmentation, Advanced Statistical Modelling Methodologies, and Future Research Directions. They also designed and developed the innovative model for early detection of AI-generated fake news presented in the case study. Alessio Jacona contributed the perspectives of journalists and press agencies and summarized the findings from the SOLARIS Use Case in Rome. All authors reviewed and approved the final version of the chapter.

References

Acemoglu, D., Ozdaglar, A., & Siderius, J. (2025). *AI and social media: A political economy perspective* (No. w33892). National Bureau of Economic Research. https://www.nber.org/papers/w33892

Ali Adeeb, R., & Mirhoseini, M. (2023). The impact of effect on the perception of fake news on social media: A systematic review. *Social Sciences, 12*(12), 674.

Al-Khazraji, S. H., Saleh, H. H., Khalid, A. I., & Mishkhal, I. A. (2023). Impact of deepfake technology on social media: Detection, misinformation and societal implications. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics, 23*, 429–441.

Arribas, C. M., Arcos, R., & Gertrudix, M. (2025). Rethinking education and training to counter AI-enhanced disinformation and information manipulations in Europe: A Delphi study. *Cogent Social Sciences, 11*(1), 2501759. https://doi.org/10.1080/23311886.2025.2501759

Bhatnagar, A., & Ghose, S. (2004). A latent class segmentation analysis of e-shoppers. *Journal of Business Research, 57*(7), 758–767.

Birke, D. (2009). The economics of networks: A survey of the empirical literature. *Journal of Economic Surveys, 23*(4), 762–793. https://doi.org/10.1111/j.1467-6419.2009.00578.x

Cazes, S. (2023). Social dialogue and collective bargaining in the age of artificial intelligence. *OECD Employment Outlook*, 221.

Chatfield, C., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2001). A new look at models for exponential smoothing. *Journal of the Royal Statistical Society: Series D (The Statistician), 50*(2), 147–159. https://doi.org/10.1111/1467-9884.00267

Corrêa, N. K., & Oliveira, N. de. (2021). Good AI for the present of humanity: Democratizing AI governance. *AI Ethics Journal, 2*(2). https://doi.org/10.47289/AIEJ20210716-2

Economides, N. (1996). The economics of networks. *International Journal of Industrial Organization, 14*(6), 673–699.

Fenga, L., & Biazzo, L. (2025). *A hybrid statistical framework for early detection of fake news.*

Forja-Pena, T., García-Orosa, B., & López-García, X. (2024). A shift amid the transition: Towards smarter, more resilient digital journalism in the age of AI and disinformation. *Social Sciences, 13*(8), 403.

Gotfredsen, S. G., & Dowling, K. (2024). *Meta is getting rid of CrowdTangle – and its replacement isn't as transparent or accessible. Columbia Journalism Review.* Retrieved October 11, 2025, from https://www.cjr.org/tow_center/meta-is-getting-rid-of-crowdtangle.php

Hermann, E. (2022). Artificial intelligence and mass personalization of communication content – An ethical and literacy perspective. *New Media & Society, 24*(5), 1258–1277. https://doi.org/10.1177/14614448211022702

Hesterberg, T. (2011). Bootstrap. *WIREs Computational Statistics, 3*(6), 497–526. https://doi.org/10.1002/wics.182

Kandlhofer, M., & Steinbauer, G. (2018). A driving license for intelligent systems. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1). https://ojs.aaai.org/index.php/AAAI/article/view/11399

Kang, L., Ye, S., Jing, K., Fan, Y., Chen, Q., Zhang, N., & Zhang, B. (2020). A segmented logistic regression approach to evaluating change in caesarean section rate with reform of birth planning policy in two regions in China from 2012 to 2016. *Risk Management and Healthcare Policy, 13*, 245–253. https://doi.org/10.2147/RMHP.S230923

Khutsishvili, K. (2024). From a smart city to wise citizens: Smart empowering with artificial intelligence. In *Smart cities to smart societies* (pp. 51–63).

Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003439325-5

Krämer, C., & Cazes, S. (2022). *Shaping the transition: Artificial intelligence and social dialogue* (OECD Social, Employment, and Migration Working Papers No. 279).

Lazar, L., & Pop, M.-I. (2021). Impact of celebrity endorsement and breaking news effect on the attention of consumers. *Studia Universitatis, Vasile Goldis, Arad – Economics Series, 31*(3), 60–74. https://doi.org/10.2478/sues-2021-0014

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Levine, S. S., & Jain, D. (2023). How network effects make AI smarter. *SSRN Electronic Journal.*https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5281829

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

Lythreatis, S., Singh, S. K., & El-Kassar, A.-N. (2022). The digital divide: A review and future research agenda. *Technological Forecasting and Social Change, 175*, 121359.

Mena, P., Barbe, D., & Chan-Olmsted, S. (2020). Misinformation on Instagram: The impact of trusted endorsements on message credibility. *Social Media + Society, 6*(2), 2056305120935102. https://doi.org/10.1177/2056305120935102

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management, 58*(3), 103434.

Mohammad, S. M., & Turney, P. D. (2013). *NRC emotion lexicon.* National Research Council Canada.

Observatory, S. I. (2022). *Memes, magnets and microchips: Narrative dynamics around COVID-19 vaccines.* Virality Project.

Outwater, M. L., Castleberry, S., Shiftan, Y., Ben-Akiva, M., Zhou, Y. S., & Kuppam, A. (2003). Attitudinal market segmentation approach to mode choice and ridership forecasting: Structural equation modeling. *Transportation Research Record, 1854*(1), 32–42. https://doi.org/10.3141/1854-04

Painter, R. W. (2023). Deepfake 2024: Will *Citizens United* and artificial intelligence together destroy representative democracy? *Journal of National Security Law & Policy, 14*, 121.

Panno, A., Pellegrini, V., De Cristofaro, V., & Donati, M. A. (2023). A measure of positive and negative perception of migration: Development and psychometric properties of the Positive and Negative Perception of Immigrants Scale (PANPIS). *Analyses of Social Issues and Public Policy, 23*(1), 73–105. https://doi.org/10.1111/asap.12338

Pawelec, M. (2022). Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital Society, 1*(2), 19.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50.

Sæbø, H. V., Ragnarsøn, R., & Westvold, T. (2020). Official statistics as a safeguard against fake news. *Statistical Journal of the IAOS, 36*(2), 435–442. https://doi.org/10.3233/SJI-190563

Shen, X.-L., & Wu, Y. (2024). Multidimensional information literacy and fact-checking behavior: A person-centered approach using latent profile analysis. In I. Sserwanga et al. (Eds.), *Wisdom, well-being, win-win* (Vol. 14597, pp. 280–297). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57860-1_20

Shete, A., Soni, H., Sajnani, Z., & Shete, A. (2021). Fake news detection using natural language processing and logistic regression. *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, 136–140. https://ieeexplore.ieee.org/abstract/document/9563292/

Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. In *Time series analysis and its applications* (pp. 75–163). Springer. https://doi.org/10.1007/978-3-319-52452-8_3

Sohrawardi, S. J., Seng, S., Chintha, A., Thai, B., Hickerson, A., Ptucha, R., & Wright, M. (2020). Defaking deepfakes: Understanding journalists' needs for deepfake detection. *Proceedings of the Computation + Journalism 2020 Conference*, 21. https://www.usenix.org/system/files/soups2020_poster_sohrawardi.pdf

Stephen, G. (2025). *Investigation and prevention of cybercrimes using artificial intelligence*.https://www.theseus.fi/handle/10024/891045

Tahat, K., Mansoori, A., Tahat, D. N., Habes, M., Alfaisal, R., Khadragy, S., & Salloum, S. A. (2022). Detecting fake news during the COVID-19 pandemic: A SEM-ML approach. *Computers, Integrated Manufacturing Systems, 28*(12), 1554–1571.

Verma, J. P. (2013). Cluster analysis: For segmenting the population. In *Data analysis in management with SPSS software*(pp. 317–358). Springer India. https://doi.org/10.1007/978-81-322-0786-3_10

Wang, J., Lu, S., Wang, S.-H., & Zhang, Y.-D. (2022). A review on extreme learning machine. *Multimedia Tools and Applications, 81*(29), 41611–41660. https://doi.org/10.1007/s11042-021-11007-7

Yan, S., Kwan, Y. H., Tan, C. S., Thumboo, J., & Low, L. L. (2018). A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Medical Research Methodology, 18*(1), 121. https://doi.org/10.1186/s12874-018-0584-9

Zaken, M. van A. (2024, January 17). *Government-wide vision on generative AI of the Netherlands* [Parliamentary document]. Ministerie van Algemene Zaken. https://www.government.nl/documents/parliamentary-documents/2024/01/17/government-wide-vision-on-generative-ai-of-the-netherlands