Zoon Politikon 16 / 2025
http://www.ejournals.eu/Zoon-Politikon/
DOI: https://doi.org/10.4467/2543408XZOP.25.001.22813

Digital disinformation and democratic destabilization: mechanisms, risks and prospects for counteracting it

Stefano Lovi

University of International Studies of Rome UNINT, Rome, Italy

e-mail: stefano.lovi@unint.eu

https://orcid.org/0009-0009-9978-8689

Abstract

In an era in which digital technologies profoundly shape access to and dissemination of information, disinformation is one of the most insidious threats to democratic societies. This contribution aims to analyse the contemporary mechanisms of creation and dissemination of disinformation online, with a particular focus on the risks arising from social media, the current functioning of their algorithms and deepfake technology. The article examines the strategies through which state and non-state actors use manipulated content to shape public opinion, erode trust in institutions, polarize public discourse and influence electoral processes. Through an interdisciplinary approach, the text offers an overview of the new forms of digital propaganda and their historical evolution, underlining the continuity between ancient practices of manipulation and modern disinformation technologies. Particular attention is paid to the concept of parasocial opinion leaders, a central figure in new digital ecosystems, and to the role of algorithms in reinforcing cognitive biases and radicalization phenomena. Finally, some reflections are offered on tools and strategies to combat disinformation, from the regulation of platforms to the strengthening of media education, with the aim of promoting critical and resilient citizenship. The article thus intends to contribute to the academic debate on disinformation as a geopolitical, social and epistemological challenge, proposing a theoretical framework useful for further research and effective policies.

Keywords: *disinformation, digital media, deepfake, propaganda, democratic processes, media education*

Date of receipt of the text by the editor: 23 April 2025;
date of acceptance for publication: 19 November 2025.

Introduction

We are living in a period of profound, and equally rapid, transformation in the way information is produced, distributed and received. The advent of digital technologies has radically reconfigured the media ecosystem, multiplying communication channels, breaking down barriers to access to information and redefining the status quo between producers and consumers of content.

While on the one hand these technologies have democratized the production of information, giving voice to subjects previously excluded from public discourse, on the other they have favoured the uncontrolled proliferation of false, misleading or deliberately manipulative content. The result is a chaotic, overflowing and sometimes toxic information environment, where the truthfulness of content is often subordinated to its ability to capture attention and generate engagement.

The contemporary information crisis manifests itself in many forms: from disinformation, which consists of false information intentionally spread to deceive, to misinformation, which unlike the previous ones is shared without malicious intent, to malinformation, in which true facts are spread in a harmful or decontextualized way. This “information disorder” represents a systemic challenge that undermines epistemic trust and, with it, the very foundations of democratic discourse (Wardle & Derakhshan, 2017). Public institutions, traditional media and informed citizens are thus exposed to new vulnerabilities, in a context in which the boundary between facts, opinions and propaganda becomes increasingly blurred.

In parallel, the communicative dynamics of disinformation have evolved. Traditional propaganda techniques, which historically presupposed a centralized control of the media, have been replaced, or enhanced, by networked practices of persuasion powered by digital infrastructures that operate in real time. The actors of contemporary disinformation are no longer just authoritarian governments or intelligence agencies, but also include extremist political groups, marketing companies, troll farms, influencers, artificial intelligences and simple social media users. Anonymity, virality and automation allow for a widespread and, often, invisible dissemination of toxic narratives, which leverage cognitive biases and emotional dynamics to spread rapidly and take root in the consciousness of the audience to whom they are addressed.

This article aims to offer an interdisciplinary and systemic analysis of disinformation in the digital age, exploring its peculiar characteristics and tracing the historical continuities with historical propaganda, which has always existed. Starting from a theoretical survey of the concepts of information, truth and trust, the cognitive mechanisms that make individuals vulnerable to disinformation, as well as the algorithmic logics that favour it, will be analysed. Contemporary media architectures, particularly those of social media, will be examined as infrastructures of cognitive manipulation, and the role of technological platforms in the construction of the public agenda will be investigated. Finally, the article will reflect on the socio-political implications of the phenomenon, questioning how disinformation contributes to the polarization of public opinion, the erosion of trust in democratic institutions and the destabilization of deliberative processes. Some intervention strategies will be outlined, both at the technological level, such as the design of more transparent and accountable algorithms, and at the cultural and educational level, with media literacy programs and the promotion of critical thinking. The aim is to strengthen democratic resilience in an era marked by algorithmic persuasion and the growing opacity of online information.

Definition and historical context of disinformation

Disinformation and propaganda are not new phenomena. They are tools that have been used strategically throughout history to serve political, religious and economic interests. Already in ancient times, rulers and political authorities have used manipulated information to strengthen their power or discredit opponents. One of the most famous examples dates back to the 4th century with the Donation of Constantine, a falsified document that granted the papacy control over vast territories. Although the forgery was discovered in the 15th century by the humanist Lorenzo Valla, the document was used for centuries for political and religious purposes (Tikkanen, 2002).

Throughout history, disinformation has taken many curious and creative forms. In the 20th century, disinformation became a powerful weapon in global conflicts, particularly during World War II. Both the Allies and the Axis Powers employed propaganda and deceptive strategies to manipulate public opinion and mislead the enemy. A striking example is Operation Fortitude, a crucial part of the D-Day strategy. The Allies created a fake army led by General Patton, complete with inflatable tanks and fake radio transmissions, to convince the Nazis that the invasion would take place at Pas-de-Calais instead of Normandy. This ruse played a significant role in the success of the Normandy landings (Downing, 2024).

From the “acta diurna” of ancient Rome to the propaganda campaigns of World War II, leaders and governments have consistently understood the power of storytelling in shaping public consciousness (Jowett & O’Donnell, 2018). However, the contemporary digital environment represents a far-reaching paradigm shift, both in the speed and personalization of disinformation.

What distinguishes today’s disinformation from its past manifestations is not only the content, but above all the infrastructure that enables its spread. The emergence of social media platforms such as Facebook, Twitter (now X), Instagram, and TikTok has created a decentralized information ecosystem in which users are both producers and consumers of content. The erosion of editorial gatekeeping and the triumph of the logic of virality over accuracy have dramatically changed the way information is validated and considered reliable (Lazer et al., 2018).

Digital disinformation differs from traditional propaganda primarily in who is practicing it, and consequently the tools used. While propaganda was typically state-driven and one-way, and therefore implemented through official or establishment channels, today disinformation can be crowd-sourced, bottom-up, and much more participatory, blurring the lines between propagandist and user.

The shift from print to the internet, and subsequently to mobile communication and algorithmic filtering, has profoundly affected the way societies communicate. The digital revolution has democratized content creation, but it has also introduced deep vulnerabilities into the public sphere. The collapse of the traditional information ecosystem, combined with the rise of influencers and alternative media sources, has created, as we have already said, fertile ground for the spread of false or misleading information (McIntyre, 2018).

Understanding the contemporary phenomenon of digital disinformation requires situating it within the broader historical and technological evolution of the Internet. The conceptual transition from Web 1.0 to the emergent paradigms of Web 4.0 and Web 5.0 has not only transformed communication practices but has also progressively expanded the capabilities, velocity, and sophistication of manipulative information campaigns. Each generational shift of the Web introduced new affordances that reshaped the relationship between users, platforms, and information flows, altering both the structure and the epistemic vulnerabilities of the digital public sphere.

The earliest phase, commonly referred to as Web 1.0, was characterized by a static and unidirectional architecture in which users primarily consumed information rather than producing it. Content was largely generated by institutions, companies, and centralized media actors, and although disinformation was certainly present, its dissemination mirrored traditional propaganda models: hierarchical, relatively slow, and dependent on the authority of established gatekeepers. As Sarowski (2017) observes, communication spaces in this early stage were limited in interactivity, curtailing the participatory dynamics that would later become typical of digital culture. Manipulation existed, but its scale was moderate, and its techniques resembled the analogy forms that had long been part of political and media history.

With the rise of Web 2.0, the cultural and technological landscape of the Internet changed radically. The emergence of social networks, blogging platforms, content-sharing sites and interactive forums inaugurated a participatory model where users became central actors in the production and circulation of content. This shift effectively dismantled the traditional hierarchy between sender and receiver: the Web transitioned from being a “read-only” space to a “read-write” environment in which millions of individuals could create, remix, and disseminate information without mediation. The consequences for disinformation were profound.

Rumours, conspiracy theories and fabricated narratives could now propagate horizontally, virally, and globally. Rani, Das and Bhardwaj (2022) note that this transition fundamentally altered the dynamics of rumour diffusion, enabling the emergence of large-scale, self-organized disinformation ecosystems that could no longer be controlled (or even fully understood) through traditional media-centric frameworks. The democratization of content production did not automatically translate into a democratization of informational quality; instead, it fostered a chaotic environment where emotionality, sensationalism and ideological tribalism often supplanted verification and epistemic rigor.

The subsequent emergence of Web 3.0 introduced further complexity. Often associated with the rise of the semantic web, big data and machine learning, this stage brought about a more personalized and algorithmically curated experience of the online world. Information was no longer presented uniformly but was selectively organized according to individual profiles, preferences, and behavioural patterns. This personalization, although beneficial from the standpoint of usability, created fertile ground for manipulation. Algorithmic systems, trained to maximize engagement rather than guarantee informational accuracy, began to shape the visibility of content in ways that intensified confirmation biases, strengthened echo chambers and facilitated micro-targeted persuasion.

The semantic web provided the technical infrastructure for sophisticated profiling, which in turn enabled the design of tailored disinformation campaigns capable of exploiting psychological vulnerabilities with unprecedented precision. In this period, manipulation shifted from a primarily collective to a deeply individualized phenomenon, where each user could be targeted with bespoke narratives designed to appeal to their fears, values and identities.

The transition toward Web 4.0 (sometimes described as the "intelligent web") further amplified these dynamics. With the incorporation of advanced artificial intelligence systems, ubiquitous connectivity and increasingly autonomous digital agents, disinformation entered a phase marked by automation, real-time adaptability and synthetic media production. AI systems no longer served merely as filters or recommendation engines; they became active producers of content capable of generating persuasive texts, realistic images, and increasingly sophisticated deepfakes. Bontridder and Poullet (2021) underline that the involvement of AI in the informational ecosystem introduces a structural duality: artificial intelligence is simultaneously a tool for detecting disinformation and one of the most powerful generators of deceptive content.

Benzie and Montasari (2022), examining AI within national-security contexts, emphasize the growing threat posed by autonomous or semi-autonomous agents capable of disseminating manipulative narratives at scale, adjusting their strategies in reaction to user behaviour and exploiting social divisions with a level of efficiency unattainable by human operators alone. In this stage, the ability to manufacture synthetic realities, audio, video and text indistinguishable from authentic content, significantly erodes the epistemic stability of public discourse, contributing to what scholars have described as the “crisis of authenticity”.

Looking toward the emergent and still largely conceptual paradigm of Web 5.0, the potential risks appear even more profound. Web 5.0 envisions a hybrid environment that combines the semantic intelligence of Web 3.0 with the emotional and immersive dimensions enabled by advances in affective computing, virtual reality, and augmented reality. Human–machine interaction becomes increasingly embodied, emotional and continuous. Within such an environment, disinformation may evolve from the dissemination of misleading textual or visual content to the creation of entire immersive experiences capable of simulating events, interactions and environments that never existed.

Manipulation becomes experiential rather than merely informational, blurring the boundaries between perception and representation. Boćkowski et al. (2022), discussing the emerging field of social cybersecurity, argue that this new stage will require a reconceptualization of how societies protect themselves from manipulation, as future disinformation is likely to exploit emotional, perceptual and even neurocognitive dimensions of user engagement. In Web 5.0, the persuasive power of mediated content is not only amplified; it becomes intimately entangled with the user’s sensory and affective experience, rendering traditional countermeasures increasingly inadequate.

Across these successive stages, the evolution of the Web reveals a clear trajectory: disinformation has become progressively more decentralized, participatory, personalized, automated and immersive. The transformation of technological infrastructures has influenced not only how disinformation is produced and disseminated but also how it is perceived, internalized and contested by users. Understanding this progression is therefore essential for conceptualizing the contemporary information disorder and for designing strategies capable of addressing the challenges posed by the next generation of manipulative practices.

This participatory culture allows anyone with access to the internet to contribute to the public narrative. The hashtag launched by Elon Musk #youarethemedianow has become emblematic in this regard. While on the one hand this opens up new opportunities for marginalized voices, on the other it facilitates the circulation of hoaxes, conspiracy theories and manipulative content. Furthermore, the enormous abundance of information leads to an “information overload”, which weakens the ability of individuals to identify reliable sources (Bawden & Robinson, 2009).

Technological innovations such as bots, artificial intelligence, and deep learning have further intensified the threat of disinformation. These tools enable the mass production of false content that may appear authentic at first glance, which is particularly dangerous in contexts such as elections, public health, or international scenarios.

Actors and Strategies of Contemporary Disinformation

Now on the practical side, disinformation campaigns are orchestrated by a wide range of actors, including state governments, political groups, ideologically motivated organizations, and individual users. State actors often exploit disinformation to influence the geopolitical landscape in order to gain internal control, as demonstrated by the work of the Russian Internet Research Agency (IRA), which played a key role in influencing public opinion during the 2016 US elections (DiResta et al., 2019).

At the same time, non-state actors such as activist groups or government-funded APT groups can use disinformation to gain economic benefits, promote ideologies, or create social unrest. Troll farms and bots, often automated or semi-automated, amplify messages to simulate popular movements or create a semblance of consensus for certain narratives, a phenomenon known as “astroturfing” (Howard & Kollanyi, 2016). The term, coined in the 1980s, denotes an unfair marketing strategy, since it often relies on people paid in a non-transparent manner to artificially produce a positive aura around the good being promoted.

In concrete terms, modern disinformation techniques include several practices, including:

1) fake news production: the deliberate creation and dissemination of false news, constructed to appear credible and attract attention. They can be published on sites specifically created for this purpose or spread through apparently authentic social profiles. During the 2016 US presidential elections, false articles declaring the death of celebrities or unfounded accusations against political candidates were spread, generating millions of shares and influencing public opinion. Usually, as in the case of the latest European elections, these are supported using *spindoctors*, i.e. accounts created specifically to send links and materials containing false news, and then quickly delete the profiles themselves once the information has entered the circulation of the platforms. In this way, they manage to hide the origin of the fake news or sources of disinformation (DiResta, 2020);
2) decontextualization: this technique consists of taking real content, such as an image, a quote, a video, and re-proposing it in a context other than the original, thus altering its meaning. A video of clashes between protesters and police in one country, shared with the caption of an event that happened elsewhere or in another historical period, can generate panic or unfounded indignation. Often, so-called CheapFakes are widely used: to better explain what they are, let's use a classic example. During the beginning of the conflict in Ukraine, a video circulated widely used to accuse Ukraine of exaggerating the number of civilian deaths in the conflict with Russia. In the incriminating video, a journalist can be seen speaking in front of the camera and behind him a myriad of bodies on the ground with a black

veil over them, but suddenly one of the bodies moves his arm to adjust the cloth that had been lifted by a gust of wind. Several accounts spread the video claiming that the Ukrainian civilian victims were actors paid to pretend to be dead. In fact, the video is from a Friday for Future protest in Vienna, where protesters lay down and put black sheets on themselves to protest climate change (Saul, 2023). The original recording titled "Wien: Demo gegen Klimapolitik"[1] was posted on YouTube before the war started;

3) memetic warfare: memes, images or short videos accompanied by ironic or satirical texts, have become powerful tools of disinformation. They can convey ideological, political or cultural messages in a simplified and viral way, often eluding normal critical filters. During election campaigns or current events, memes that ridicule one candidate or praise another can spread rapidly, influencing the public's judgment without providing in-depth analysis;
4) use of Deepfake technology: these are audiovisual contents manipulated through artificial intelligence, which allow for the extremely realistic simulation of the faces and voices of real people. These tools make it increasingly difficult to distinguish real contents from manipulated ones, raising serious concerns about the erosion of information credibility. Deepfakes have already been used in political campaigns and in revenge porn, but their potential extends to the discrediting of public figures, the manipulation of financial markets and the incitement to violence (Chesney & Citron, 2019).

The topic of deepfakes deserves further exploration, as the challenge posed by this technology is twofold. First, they are a powerful means of

---

[1] Wien: Demo gegen Klimapolitik. *OE24.TV*. YouTube. Retrieved April 16, 2025, from https://www.youtube.com/watch?v=203Rph8DcUY.

active disinformation, spreading blatant lies that appear visually authentic. Second, they contribute to what some scholars have called the "liar's dividend" (Fallis, 2021). This term refers to the situation in which people can deny the authenticity of video, image and audio evidence, especially if it is depicted inappropriate or criminal behaviour, by claiming that it was artificially generated. As the public becomes increasingly aware of new video manipulation technologies, it is likely that audiovisual evidence, even if real, will increasingly be denounced as fake. This denial tactic has already been widely used in politics, for example when former US President Donald Trump claimed that a recording of him on the television show Access Hollywood, in which he said demeaning comments about women, was fake, while the host confirmed its authenticity (Victor, 2017).

On March 16, 2022, a short 72-second video message from Volodymyr Zelenskyy, wearing his iconic olive-green shirt, appeared on the Ukraine24 website, in which the President, rather than urging Ukrainians to continue their fight, implored them to lie down, lay down their weapons and surrender.[2] Of course, the video quickly spread on VKontakte, Telegram and other social media platforms, where it was reported by global media, followed by Zelenskyy's office which immediately denied the authenticity of the video (Simonite, 2022). Nonetheless, it was the first use of such technology to produce a false statement by a president in an armed conflict (Byman et al., 2023).

The Structural Link Between Echo Chambers and Cognitive Biases

The dynamics of online disinformation cannot be understood without a careful analysis of the interaction between algorithmic architectures

[2] Deepfake video of Volodymyr Zelensky surrending surfaces on social media. *The Telegraph*. YouTube. Retrieved April 16, 2025, from https://www.youtube.com/watch?v=X17yrEV5sl4&ab_channel=TheTelegraph.

and human cognitive mechanisms. The information manipulation techniques described in the previous chapter thrive in digital ecosystems governed by the attention economy, in which metrics such as likes, shares, retweets and views define the visibility and circulation of information.

This logic of attention, rooted in precise economic objectives, namely increasing the time spent on the platform and, consequently, advertising revenues, favours content that is emotionally charged, polarizing and easily digestible. In this context, the problem lies not only in the way algorithms work, but also in the way they are optimized: not for their information value, but for the emotional reactivity they can generate (Tufekci, 2015). Such optimization leads to the systematic overexposure of sensational content to the detriment of accurate and contextualized information.

Not only algorithmic codes contribute to this dynamic, but also user behaviour. The use of digital information is often characterized by short attention spans and preferences for short, visual, and emotionally impactful content. This makes disinformative content, often constructed with manipulative purposes and therefore designed to strike quickly, more effective in terms of social metrics than traditional journalistic content, which requires critical reading and contextualization.

In this environment, algorithmic selection acts as a filter that enhances various cognitive biases already present in the user's mind. For example, confirmation bias pushes users to favour information that confirms their pre-existing beliefs, ignoring or devaluing dissonant ones (Casad & Luebering, 2025). This behaviour, often implemented unconsciously, is further reinforced by social media algorithms, which tend to propose content like the preferences previously expressed by the user, favouring a unilateral worldview where contradictory and constructive debate find little space.

The result is the formation of so-called echo chambers and filter bubbles, or digital environments in which the user is exposed almost exclusively to perspectives that reflect or reinforce his or her opinions. Such environments not only reduce exposure to dissent but also limit the ability to develop critical thinking and to question one's cognitive schemas (Spohr, 2017). Contrary information is ignored, discredited or perceived as a threat, reinforcing the dichotomy of an "Us" versus a "Them" that fuels polarization.

The implications of these dynamics are profound and affect both the cognitive and sociopolitical spheres. From a psychological point of view, repeated exposure to ideologically coherent content strengthens group identity and increases trust in the narratives shared by one's ingroup. This identity strengthening is a powerful driver of polarization, which can intensify to the point of leading to ideological radicalization. One of the most cited examples in literature concerns the functioning of YouTube's recommendation engine, which according to several studies can push users towards increasingly extreme content over time, through a sequence of progressive suggestions (Ribeiro et al., 2019). In these cases, the logic of engagement fuels a path of algorithmic radicalization, in which the user is gradually led from mainstream content to conspiracy or extremist narratives.

Furthermore, a mechanism that is greatly amplified by social platforms is the so-called bandwagon effect. This is a cognitive and social bias that describes the tendency of individuals to adopt beliefs, behaviours or opinions because they perceive that they are already accepted or followed by the majority. In other words, the more people seem to believe in something, the more likely it is that others will join the consensus, often without independent critical evaluation. This psychological mechanism, well documented in social psychology and collective behaviour models, plays a fundamental role in the propagation of disinformation on social media (Cialdini & Goldstein, 2004).

In the contemporary digital ecosystem, the bandwagon effect is amplified by the algorithmic dynamics of social platforms. Users are constantly exposed to popularity metrics, represented by the number of likes, shares, comments and views, which act as social signals and implicitly suggest the "truth" or legitimacy of a content. This can lead to the normalization of false or misleading narratives, especially when such content goes viral or is associated with influential public figures. In this context, the plausibility of a piece of information depends not so much on its adherence to facts as on its perceived social acceptance.

For example, during the COVID-19 pandemic, many conspiracy theories, such as the idea that the virus was deliberately created in a laboratory or that vaccines contained microchips, gained traction not because they were supported by scientific evidence, but because they were amplified by thousands of online interactions. The growing perceived consensus induced other users to adhere to these beliefs, strengthening a self-referential circuit of social validation. This cycle is further fueled when disinformation is supported by tight-knit digital communities, where dissent is discouraged and collective buy-in is rewarded.

From the perspective of digital platforms, the bandwagon effect is also exploited commercially. Content that shows signs of popularity is boosted by recommendation algorithms, as it is more likely to generate engagement. This creates a positive feedback effect, where what is popular becomes even more visible, regardless of its informational accuracy. This mechanism makes it extremely difficult to stop the viral spread of disinformation once it has reached a critical threshold of visibility.

An in-depth analysis of echo chambers and cognitive biases would be incomplete without addressing their central role in the emergence, consolidation and diffusion of conspiracy theories. The COVID-19 pandemic represented a turning point in the global information ecosystem, serving as an unprecedented laboratory for observing how crises of uncertainty, algorithmic architectures and psychological predispositions interact to produce large-scale conspiratorial thinking. During this period, the dynamics of online disinformation became especially intertwined with the human need for causal explanations, emotional reassurance and group belonging, factors that greatly amplified susceptibility to conspiracy narratives.

The development of conspiracy beliefs has been shown to correlate strongly with specific cognitive dispositions, especially low analytical thinking, high intuition-driven reasoning and elevated levels of epistemic mistrust. Vital et al. (2025) demonstrate that when individuals search for COVID-19–related information online, those with weaker analytical reasoning capacities are more likely to interpret ambiguous or contradictory information through a conspiratorial lens. Their study highlights that the digital information environment does not merely expose individuals to conspiracy content; it actively interacts with their cognitive style. Search engines and social media, optimized for personalization, tend to reinforce the interpretative frameworks already preferred by the user. As a result, users who engage in less reflective thinking often find themselves enclosed in informational bubbles where conspiratorial content is both more salient and more cognitively fluent than scientifically grounded explanations.

At the same time, research by Duplaga (2020) shows that conspiracy beliefs related to COVID-19 emerged in clear association with sociodemographic factors, institutional mistrust and perceived lack of control. The pandemic, characterized by rapid decision-making, evolving scientific knowledge and the constant presence of uncertainty, created an ideal backdrop for the proliferation of conspiratorial narratives. According to Duplaga, individuals exhibiting lower trust in health institutions and government authorities were significantly more vulnerable to disinformation, often treating conspiracy explanations not simply as alternative theories but as coherent narratives that offered psychological comfort during a period of collective anxiety. The spread of these beliefs was thus not only a consequence of exposure to misleading content, but also an outcome of a deeper sociopolitical and epistemic crisis.

Echo chambers played a decisive role in reinforcing these vulnerabilities. Digital platforms, driven by recommendation algorithms that privilege engagement and emotional intensity over informational accuracy, created homogeneous environments where conspiracy theories circulated without significant friction. Within these closed networks, misinformation was continuously validated through social reinforcement, while scientific or corrective information was dismissed as manipulative or politically motivated. The interpretative isolation fostered by echo chambers also encouraged the emergence of strong group identities, often built around narratives of resistance to perceived elites, experts or institutions. These identities provided an additional psychological incentive to accept conspiratorial worldviews, as belonging to the group offered emotional security, social validation and a clear sense of distinction between “us” and “them”.

The pandemic further exemplified how conspiracy theories operate as self-sealing belief systems. Once internalized, conspiracy frameworks reinterpret any contradictory evidence as proof of the conspiracy itself, a dynamic that makes debunking strategies far less effective. In online communities shaped by echo chambers, users frequently developed meta-conspiratorial beliefs, such as the idea that corrections issued by health authorities were themselves part of a coordinated cover-up. This recursive logic was reinforced by cognitive biases already intensified in digital settings, including confirmation bias, proportionality bias and the illusory truth effect. The latter, in particular, became especially relevant: repeated encounters with the same false claim across multiple online spaces increased the likelihood that users would eventually consider it true, regardless of prior skepticism.

The COVID-19 crisis therefore illustrates how conspiracy theories arise not simply from ignorance or irrationality, but from the complex interplay between cognitive predispositions, emotional needs, structural features of digital platforms and sociopolitical contexts marked by high polarization and institutional fragility. In this sense, conspiracy thinking should be understood as an emergent property of the contemporary information environment — a systemic outcome of the interaction between human psychology and algorithmic infrastructures.

Integrating these insights into the broader discussion of echo chambers and cognitive biases underscores the need to conceptualize disinformation not only as the dissemination of false content but as a form of epistemic engineering that operates through the reconfiguration of interpretative habits, trust networks and emotional narratives. The pandemic made clear that once conspiratorial frameworks take root within algorithmically reinforced communities, they can profoundly influence political behaviour, public health compliance, interpersonal trust and the overall stability of democratic discourse. Addressing this challenge requires a multi-layered approach that takes into account the psychological, technological and social dimensions of contemporary disinformation, recognizing conspiracy thinking as one of its most consequential and entrenched manifestations.

The Role of Parasocial Leaders

The transformation of the public sphere in the digital age has given rise to a new type of opinion leader: the social media influencer. Unlike traditional opinion leaders, who were typically pundits, intellectuals, or journalists whose authority derived from institutional roles or recognized expertise, influencers often draw their power from visibility and charisma rather than knowledge or analytical expertise. This shift coincides with the emergence of what Horton and Wohl famously called the "parasocial relationship" way back in 1956, a one-sided psychological bond that media consumers form with mediated figures who are unaware of their existence.

In the contemporary digital ecosystem, *parasociality* has been dramatically amplified by constant interactivity and algorithmic personalization. Influencers present themselves in intimate and accessible ways, sharing personal stories, behind-the-scenes content, and daily routines, thereby encouraging audiences to develop feelings of trust, familiarity, and emotional closeness. These pseudo-relationships strengthen the persuasive power of influencers, who are often perceived as more authentic and trustworthy than traditional media figures (Abidin, 2016). In the contemporary digital ecosystem, this type of relationship has not only become normalized, but has been amplified and made more sophisticated.

The result is the creation of an “algorithmic intimacy” (Bishop, 2021), in which emotional engagement is mediated but perceived as authentic. Users feel connected to influencers as true friends or confidants, which creates fertile ground for the suspension of critical thinking. Unlike traditional media, where authority was linked to journalistic standards and verification mechanisms, parasocial communication is based on an affective and performative dynamic, which rewards narrative coherence, the *spectacularization* of personal experience, and charisma.

Building online authority depends less on institutional credentials and more on branding strategies, narrative coherence, and audience engagement. Influencers cultivate a loyal following by creating a distinctive persona and maintaining ongoing, responsive interactions with their audiences. Their power does not necessarily lie in offering factual or well-reasoned commentary, but in their ability to capture attention, resonate emotionally, and cultivate perceived intimacy. This results in a communication environment where emotional connection and entertainment value often outweigh rational reflection (Marwick & Lewis, 2017).

A case in point are wellness or alternative health influencers, some of whom have played a significant role in spreading conspiracy theories during the COVID-19 pandemic, promoting pseudoscientific information about vaccines, unapproved therapies, and alleged global conspiracies. The bond of trust established with followers, combined with an emotional and "anti-establishment" narrative, has made these actors incredibly effective in conveying disinformation messages, often with concrete effects on public health (Cinelli et al., 2020).

These dynamics have profound implications for the formation of public opinion. The rise of parasocial leaders facilitates the viral spread of unverified and emotionally charged content that reinforces tribal divisions rather than promotes democratic discourse. When influencers intervene in political or ideological debates, their messages often bypass traditional mechanisms of scrutiny and validation. Their followers, emotionally invested in the influencer's personality, are more likely to accept their claims uncritically, contributing to the normalization and amplification of misinformation or polarizing narratives (Tandoc, Lim & Ling, 2019).

Democratic Consequences of Disinformation

One of the most alarming effects of disinformation in the digital age is the erosion of trust in democratic institutions. By polluting the information environment with false or misleading narratives, disinformation undermines trust in the media, the scientific community, and public governance. This creates what scholars call an "epistemic crisis", in which the very idea of shared truth becomes suspect and consensual reality shatters into partisan echo chambers (Lewandowsky, Ecker & Cook, 2017).

As users increasingly struggle to distinguish between facts and fiction, the legitimacy of democratic institutions suffers. Trust in electoral processes, judicial decisions and public policies becomes fragile, especially when malicious actors intentionally sow doubts about the integrity of these systems. This is particularly evident in moments of crisis, such as the COVID-19 pandemic or the recent highly contentious elections in Romania, when disinformation can fuel conspiracy theories, public confusion and institutional paralysis.

Disinformation also plays a central role in intensifying social polarization and conflict. By promoting divisive and emotionally provocative narratives, disinformation deepens ideological fractures and consolidates group identities. Digital platforms, designed to maximize engagement through emotionally resonant content, tend to privilege outrage, fear and resentment: emotions that are highly effective in driving user interaction but toxic to deliberative democracy (Bail et al., 2018).

This polarized environment fuels the rise of extremism and populist rhetoric, often making compromise or dialogue impossible. Public forums, both online and offline, become battlegrounds between opposing realities, where trust and empathy are replaced by hostility and suspicion. These conditions are ripe for manipulation by actors seeking to destabilize social cohesion or exploit divisions for political gain.

Perhaps most significantly, disinformation has been shown to have a direct impact on electoral processes. The 2016 US presidential election and the Brexit referendum are stark examples of how digital disinformation campaigns, often coordinated by foreign entities, can distort public opinion, limit voter turnout, or distort public debate. These interventions, amplified by micro-targeted ads and bot-generated content, can subtly influence outcomes without leaving obvious traces, thus evading accountability (Bennett & Livingston, 2018).

And democracies are having a hard time curbing the problem. The Georgescu case, which required the intervention of Romania's Constitutional Court following the vote count, is emblematic. The resulting street protests, staged not only by Georgescu's supporters, illustrate this systemic weakness. The Central Election Office's March 4 decision to invalidate his candidacy sparked violent protests in Bucharest, with Georgescu supporters clashing with police. The decision to annul November's election was unprecedented and handed a propaganda victory to Moscow, which has long dismissed Western democracy as a farce in which elites rig results to suppress unorthodox views and ensure an outcome favourable to the liberal establishment.

Strategies to Counter Disinformation

Countering disinformation in all its forms is one of the most urgent and complex challenges for contemporary democracies today. The polymorphic and adaptive nature of the phenomenon, capable of evolving extremely rapidly in response to new technologies, global events and changes in user behaviour, requires a global, multilayered and deeply coordinated approach. There is no single solution: however, several researchers argue that effective responses must intertwine regulation, media literacy, technological innovation, interdisciplinary research and civic engagement (Rød, Pursiainen & Eklund, 2025; McDougall, 2019).

One of the most debated areas of intervention in recent years concerns the regulatory supervision of digital platforms. Long self-regulated according to market logic, big tech is now at the center of public and political debate for the role they play in the spread of harmful content, including fake news, hate speech and coordinated manipulation campaigns. Governments, supranational authorities and advocacy bodies are calling for greater transparency in the management of algorithms, content moderation and use of personal data. A prime example is the European Union's Digital Services Act (DSA), which came into force in 2023, which imposes stringent obligations on large platforms to assess and mitigate the "systemic risks" associated with their services, such as electoral disinformation, radicalization or manipulation through bots and automated accounts (Gorwa, 2019).

However, regulatory regulation alone is not enough. Laws can set external limits and establish oversight mechanisms, but they do not act directly on the cognitive and emotional vulnerabilities of individuals, which are one of the main vectors of disinformation dissemination. For this reason, an effective counteraction strategy must necessarily include interventions aimed at strengthening media and digital literacy. Educating citizens from school age to understand the logic of the information system, to recognize reliable sources, to distinguish between opinion and verified data, and to develop critical thinking towards online content is an essential investment for democratic resilience in the long term. In this context, innovative educational initiatives such as prebunking programs have shown promising results.

Research by various authors has shown that brief interventions, such as videos or interactive games that expose users to rhetorical techniques common in disinformation, such as scapegoating, exaggerated emotionality or the use of false experts, can significantly increase individuals' critical capacity (Linden van der, Roozenbeek & Compton, 2020). Unlike traditional debunking, which seeks to correct false information after it has been internalized, pre-bunking acts ex ante, strengthening "cognitive defences" against future attempts at manipulation.

At the same time, a collective and multilateral effort involving a broad spectrum of social actors is needed. Countermeasures against disinformation cannot be entrusted exclusively to states or technology companies. Structured collaboration between public institutions, digital platforms, university researchers, journalists, educators and civil society organisations is needed. Only through an integrated and collaborative approach is it possible to address the complexity of the current information challenge. Initiatives such as international fact-checking consortia, such as the International Fact-Checking Network, interdisciplinary teams for the analysis of electoral data, or task forces against foreign influence, such as the Center for Strategic Communication of the European Commission, represent models of virtuous cooperation that can be replicated and strengthened.

Finally, it is essential to invest in the development of responsible technologies, capable of detecting and containing disinformation in real time, without falling back into mechanisms of censorship or arbitrary control. Artificial intelligence can play a positive role in this context, for example in the automatic detection of deepfakes, in the flagging of suspicious content or in the promotion of reliable content. However, it is essential that these technologies are designed with transparency and human oversight criteria, to prevent them from becoming tools of potential abuse.

However, this balance can be seriously compromised if some of these players, especially the large social platforms, choose to abandon or drastically reduce the systems for moderating and verifying content (McMahon, Kleinman & Subramanian, 2025). Elon Musk, since he acquired X (formerly Twitter), has promoted a more "libertarian" vision of the platform, cutting many of the teams dedicated to moderating content and reducing the tools for verifying sources. Mark Zuckerberg, despite having adopted more cautious policies in the past, has also scaled down in some cases the automatic and manual controls on Facebook and Instagram content, pushing towards a more automated and less centralized model. On a regulatory level, the lack of cooperation on the part of the platforms makes it much more difficult to apply regulations such as the Digital Services Act, because there are no interlocutors or internal systems to collaborate with. On a technological level, however, the abandonment of algorithmic and human verification tools leaves the field open to misleading or manipulative content, which spreads rapidly thanks to the viral logic of the platforms themselves.

Furthermore, on a psychological and social level, the perception that "everyone can say what they want, without filters" fuels a climate of generalized distrust, in which it becomes more difficult to distinguish between reliable sources and fake news. In a society where platforms no longer play an active role in moderation, the burden of media literacy falls entirely on the individual, who however is not always ready or equipped to bear this burden alone.

Conclusions

Digital disinformation is neither a marginal nor transitory phenomenon: it constitutes a structural challenge for the functioning of contemporary democracies, directly affecting the quality of public debate, trust in institutions and social cohesion. Its pervasive nature, the speed with which it spreads thanks to the viral architectures of digital platforms and the intrinsic difficulty in controlling its circulation, make clear the urgency of systemic, multilevel responses based on a solid interdisciplinary approach.

As discussed in this contribution, understanding disinformation requires the integration of skills from different fields: from cognitive psychology, which helps us explain individual vulnerability to manipulative content, to data science, which analyses information flows and algorithmic logics, through communication theory, digital sociology and political science. Only an approach that can grasp the interconnections between these levels can produce effective countermeasures, capable of going beyond the simple removal of false content and addressing the structural causes of the phenomenon.

This article has shown how disinformation thrives in disintermediated communication environments, governed by logics of engagement rather than information accuracy. The algorithms that regulate the visibility of content on social media — optimized to maximize dwell time, interactions and monetization — tend to favour sensational, polarizing and emotionally charged messages, thus contributing to the formation of “echo chambers” and “filter bubbles” that reinforce pre-existing beliefs and reduce exposure to alternative views. In this context, disinformation not only spreads more easily, but also becomes more difficult to recognize, confusing itself with legitimate content and exploiting deeply rooted cognitive mechanisms such as confirmation bias.

No less relevant is the figure of the parasocial leader, a key element in the new ecology of disinformation. It combines the ability to influence psychologically, the power of media amplification and the absence of institutional constraints, operating within an algorithmic structure that rewards precisely those communicative traits, immediacy, emotion, polarization, that facilitate the spread of misleading content. In an era in which trust in media institutions is declining, parasocial influencers embody a model of post-truthful authority, capable of shaping behaviours, opinions and values on a massive scale, but outside of any systematic verification process.

However, far from falling into a pessimistic techno-determinism, it is important to underline that the same technologies that facilitate information manipulation can also be rethought and oriented towards more democratic purposes. The design of more transparent algorithms, the promotion of participatory reporting mechanisms, the support for independent journalistic production and the creation of more inclusive online deliberative spaces represents possible directions for a healthier and more pluralistic information ecosystem.

In conclusion, addressing digital disinformation does not simply mean correcting factual errors or repressing illegitimate content, but implies a profound rethinking of the cultural, technological and political conditions that regulate our experience of information. It means questioning the role of citizens as active agents in the communication process, the ethical responsibility of digital platforms and the role of institutions in ensuring free, reliable and accessible information. Only through a collective commitment, which combines critical education, regulatory innovation and epistemic solidarity, will it be possible to build a more aware digital citizenship, capable of resisting the pitfalls of manipulation and contributing to a more robust, informed and participatory democracy.

References

Abidin, C. (2016). Visibility labour: Engaging with Influencers' fashion brands and #OOTD advertorial campaigns on Instagram. *Media International Australia*, 161(1), 86-100. https://doi.org/10.1177/1329878X16665177.

Bail, CA, Argyle, LP, Brown, TW, Bumpus, JP, Chen, H., Fallin Hunzaker, MB, Lee, J., Mann, M., Merhout, F. &. Volfovsky A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115 (37) 9216-9221. https://doi.org/10.1073/pnas.1804840115.

Bawden, D., Robinson, L. (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2), 180-191. https://doi.org/10.1177/0165551508095781.

Bennett, WL, Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, 33(2), 122-139. https://doi.org/10.1177/0267323118760317.

Benzie, A., Montasari, R. (2022). Artificial Intelligence and the Spread of Mis- and Disinformation. [In:] Montasari R. (Ed.), *Artificial Intelligence and National Security*, 1-17. Cham: Springer. https://doi.org/10.1007/978-3-031-06709-9_1.

Bishop, S. (2021). Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias. *Social Media + Society*, 7(1). https://doi.org/10.1177/20563051211003066.

Boćkowski, D. et al. (2022). *Dezinformacja – inspiracja – społeczeństwo: social cybersecurity*. Białystok: Wydawnictwo Uniwersytetu w Białymstoku. http://hdl.handle.net/11320/13835.

Bontridder, N., Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, 3, e32. https://doi.org/10.1017/dap.2021.20.

Byman, D.L., Gao, C., Meserole, C., Subrahmanian, V.S. (2023). Deepfakes and International Conflict. *Foreign Politicy at Brookings*. Retrieved April 16, 2025, from https://www.brookings.edu/wp-content/uploads/2023/01/FP_20230105_deepfakes_international_conflict.pdf.

Casad, BJ, Luebering, JE (2025). confirmation bias. *Britannica*. Retrieved April 17, 2025, from https://www.britannica.com/science/confirmation-bias.

Chesney R., DK Citron (2019), Deepfakes and the New Disinformation War. *Foreign Affairs*, 98(1), 147-155. https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war.

Cialdini, RB, Goldstein, NJ (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55, 591-621. https://doi.org/10.1146/annurev.psych.55.090902.142015.

Cinelli, M. et al (2020). The COVID-19 social media infodemic. *Scientific Reports*, 10, 16598. https://doi.org/10.1038/s41598-020-73510-5.

DiResta, R. (2020). The Spin Doctors Are In. Guest: Renée DiResta. *Center for Human Technology*. Retrieved April 16, 2025, from https://www.humanetech.com/podcast/17-the-spin-doctors-are-in.

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J. & Johnson, B. (2019). The Tactics & Tropes of the Internet Research Agency. *Stanford Internet Observatory*. Retrieved April 15, 2025, from https://digitalcommons.unl.edu/sociologyfacpub/566/.

Howard, PN, Kollanyi, B. (2016). Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. *Social Science Research Network*. Retrieved April 15, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2798311.

Downing, T. (2024). D-Day deception Operation Fortitude: The World War Two army that didn't exist. *BBC*. Retrieved April 16, 2025, from https://www.bbc.com/culture/article/20240531-d-day-deception-operation-fortitude-the-world-war-two-army-that-didnt-exist.

Duplaga, M. (2020). The Determinants of Conspiracy Beliefs Related to the COVID-19 Pandemic in a Nationally Representative Sample of Internet Users. *International Journal of Environmental Research and Public Health*, 17(21), 7818. https://doi.org/10.3390/ijerph17217818.

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philos Technol*, 34(4), 623-643. doi: 10.1007/s13347-020-00419-2. https://pubmed.ncbi.nlm.nih.gov/32837868/.

Gorwa, R. (2019). What is platform governance?. *Information, Communication & Society*, 22(6), 854-871. https://doi.org/10.1080/1369118X.2019.1573914.

Horton, D., Wohl, RR (1956). Mass communication and para-social interaction. *Psychiatry: Journal for the Study of Interpersonal Processes*, 19, 215-229. https://doi.org/10.1080/00332747.1956.11023049.

Jowett, G.S., O'Donnell, V. (2018). *Propaganda & Persuasion*. Thousand Oaks, SAGE Publications.

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

Lewandowsky, S., Ecker, UKH, Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the “Post-Truth” Era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008.

Linden van der, S., Roozenbeek, J., Compton, J. (2020). Inoculating Against Fake News About COVID-19. *Front Psychol*, 23:11:566790. https://doi.org/10.3389/fpsyg.2020.566790.

Marwick, A., Lewis, R. (2017). Media manipulation and disinformation online. *Data & Society*. Retrieved April 15, 2025, from https://datasociety.net/library/media-manipulation-and-disinfo-online/.

McDougall, J. (2019). Media Literacy versus Fake News: Critical Thinking, Resilience and Civic Engagement. *Media Studies*, 10(19). https://hrcak.srce.hr/ojs/index.php/medijske-studije/article/view/8786.

McIntyre, L. (2018). Post-Truth. *European Journal of Communication*, 33(5), 574-575. https://doi.org/10.1177/0267323118799184.

McMahon, L., Kleinman, Z., Subramanian, C. (2025). Facebook and Instagram get rid of fact checkers, *BBC*. Retrieved April 16, 2025, from https://www.bbc.com/news/articles/cly74mpy8klo.

Rani, N., Das, P., Bhardwaj, A. K. (2022). Rumor, misinformation among web: A contemporary review of rumor detection techniques during different web waves. *Concurrency and Computation: Practice and Experience*, 34(1), e6479. https://doi.org/10.1002/cpe.6479.

Ribeiro M.H., Ottoni R., West R., Almeida V., Meira W. (2019). Auditing Radicalization Pathways on YouTube. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. *Association for Computing Machinery*, 131-141. https://doi.org/10.48550/arXiv.1908.08313.

Rød, B., Pursiainen, C. & Eklund, N. (2025). Combatting Disinformation – How Do We Create Resilient Societies? Literature Review and Analytical Framework. *Eur J Secur Res*. https://doi.org/10.1007/s41125-025-00105-4.

Sarowski, Ł. (2017). Od Internetu Web 1.0 do Internetu Web 4.0 – ewolucja form przestrzeni komunikacyjnych w globalnej sieci. *Rozprawy Społeczne/Social Dissertations*, 11(1), 32-39. https://doi.org/10.29316/rs.2017.4

Saul, F. (2023). Bewegung in Leichensäcken – dieser Fake-Clip wird seit Kriegsbeginn immer wieder verbreitet. *Stern*. Retrieved April 15, 2025, from https://www.stern.de/digital/webvideo/fake-or-no-fake/bewegung-in-leichensaecken--ukraine-fake-im-faktencheck--video--33235984.html.

Simonite, T. (2022). A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be. *WIRED*. Retrieved April 16, 2025, from https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/.

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3), 150-160. https://doi.org/10.1177/0266382117722446.

Tandoc, EC, Lim, D., Ling, R. (2019). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, 21(3), 381-398. https://doi.org/10.1177/1464884919868325.

Tikkanen, A. (2002). Donation of Constantine. *Britannica*. Retrieved April 16, 2025, from https://www.britannica.com/topic/Donation-of-Constantine.

Tufekci, Z. (2015). Facebook Said Its Algorithms Do Help Form Echo Chambers, and the Tech Press Missed It. *New Perspectives Quarterly*, 32(3), 9-12. https://doi.org/10.1111/npqu.11519.

Victor, D. (2017). "Access Hollywood" Reminds Trump: "The Tape Is Very Real". *The New York Times*. Retrieved April 16, 2025, from https://www.nytimes.com/2017/11/28/us/politics/donald-trump-tape.html.

Vital, N., Chevalier, A., Dosso, Ch., Trémolière, B. (2025). Conspiracy beliefs and analytical thinking in COVID-19 information web search. *Computers in Human Behavior Reports*, 20, 100804. https://doi.org/10.1016/j.chbr.2025.100804.

Wardle, C., Derakhshan, H. (2017). Information Disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*. Retrieved April 16, 2025, from https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

DOI: 10.4467/2543408XZOP.25.001.22813

About Autor

Stefano Lovi is a PhD student at the University of International Studies of Rome UNINT. Graduated from UNINT in 2022 in the degree course "Investigation, Crime and International Security", his main research topics concern conspiracy theories, disinformation and radicalization phenomena.

https://orcid.org/0009-0009-9978-8689

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International license for use and distribution (CC BY 4.0).

Jak cytować ten aktykuł / How to cite this paper:
Lovi, S. (2025). Digital disinformation and democratic destabilization: mechanisms, risks and prospects for counteracting it. *Zoon Politikon* No. 16, 1-32
https://doi.org/10.4467/2543408XZOP.25.001.22813