

A black and white photograph of a person with curly hair, seen from behind, looking at a computer screen. The screen displays a grid pattern, possibly representing data or a user interface. The person is wearing a dark jacket. The background is slightly blurred, showing what appears to be a desk or office environment. The overall image has a grainy, high-contrast aesthetic.

Misinformation, Disinformation & Content Moderation

Explore the rising challenge of combating false information online in this introduction to the tools, strategies and collaborations needed to tackle this issue.

Contents

- 02** A brief history of misinformation
- 05** What contributed to the rise of misinformation?
- 08** Birdwatching: how Twitter cultivated its counter speech approach to managing misinformation
- 10** WebPurify's frontline battles on misinformation
- 12** Future risks
- 15** Strategies and solutions
- 18** NewsGuard: fighting to improve digital literacy
- 20** Glossary of terms



RYOJI IWATA/UNSPLASH

A Brief History of Misinformation

Misinformation is defined as false or misleading information that is *unintentionally* presented as fact. It's an important distinction to make, as disinformation is when false or misleading information is deliberately shared as fact.



ABSOLUTVISION/UNSPLASH

Misinformation and disinformation have a rich historical context stretching far before the advent of digital media. A notable example from 1835 is “The Great Moon Hoax,” where a tabloid depicted fictional inhabitants on the moon, illustrating that disinformation is not a new phenomenon but has been part of media history since the invention of the printing press.

Adding to this historical context, Holocaust denial is a classic example of disinformation, where falsities were deliberately spread to question the existence of the Holocaust, despite overwhelming historical evidence. This disinformation, often rooted in anti-Semitic agendas, aims to undermine and distort historical truth for ideological purposes.

In more contemporary examples, the Sandy Hook conspiracy, popularized by conspiracy

theorist Alex Jones, initially emerged as disinformation with false claims that the school shooting was a hoax, causing immense distress to the families of victims. This led to a landmark defamation case, where Jones was ordered to pay over \$1 billion in defamation charges, marking a substantial victory against disinformation.

More recently, in the case of the Capitol riot on January 6, false claims made about the US Capitol police’s response to the protesters would be considered disinformation if those spreading the claims did so with the intention to manipulate public perception of the events. Similarly, the incident involving a shooting at a pizza restaurant in Washington D.C., spurred by fabricated stories on social media, represents the severe impact of disinformation when it is crafted and spread with the purpose of influencing or causing harm.

Disinformation also has a habit of becoming misinformation. During the Covid-19 pandemic, disinformation regarding vaccine safety emerged as a prominent challenge. Initially, this disinformation was propagated deliberately by certain groups or individuals, aiming to create distrust and fear about the vaccines' efficacy and safety. These false claims, ranging from exaggerated side effects to unfounded conspiracy theories about vaccine ingredients, were strategically crafted to undermine public confidence in the vaccination campaign.

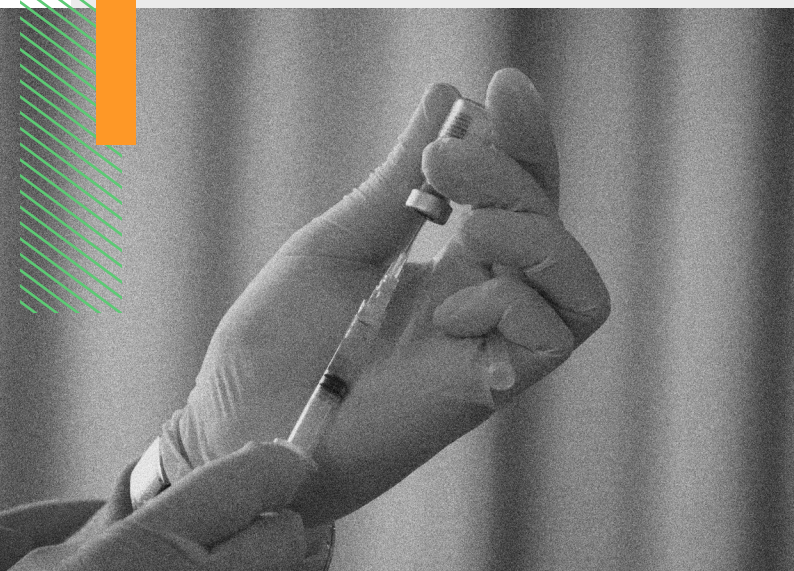
However, as this disinformation permeated social media and other communication channels, it began to morph into misinformation. Well-meaning individuals, influenced by the misleading content and often lacking access to accurate information, started to share these claims unknowingly. This transformation from disinformation to misinformation significantly impeded public health efforts, as it led to vaccine hesitancy and resistance among wider populations who were unintentionally spreading inaccuracies they believed to be true. This phenomenon

During the Covid-19 pandemic, disinformation regarding vaccine safety emerged as a prominent challenge.

underscored the complex and dynamic nature of information spread in the digital age, where distinguishing between deliberate falsehoods and unintentional misinformation has become increasingly challenging.

In each case, the key factor distinguishing misinformation from disinformation is the presence of intent to deceive behind the spread of false information.

The spread of misinformation is not just confined to political or emergency situations but extends to health misinformation, impacting public health responses, and even fostering conspiracy theories. These historical and contemporary examples underline the critical role of content moderation in curbing misinformation and upholding truth in the public domain.



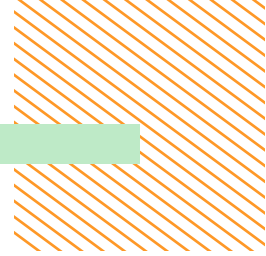
MUFID MAJUNUN/UNSPLASH



ROBIN WORRALL/UNSPASH

What contributed to the rise of misinformation?

Misinformation is not a modern problem – it’s one of the oldest tricks in the book. There are examples of it being employed throughout history in an attempt to influence people and change public opinion. But the momentum, reach and open access of today’s online platforms makes it possible to disseminate and amplify misleading information in highly efficient ways.



Being part of the conversation has never been easier. The democratization of information and the dynamic social media landscape provides myriad opportunities for user-generated content (UGC) to connect with a global audience. It's big business too. Third parties understand the value in harnessing the authentic voice of a community; the global User Generated Content Platform Market is estimated to be [worth USD \\$71.3 billion by 2032.](#)

But there is also an ugly side to UGC. Its capacity to cut through, build trust and cultivate engagement has made it a target for abuse. Influencers that are economical with the truth about product promotions are one thing, but targeted disinformation campaigns by nefarious groups are quite another.

Echo chambers, filter bubbles and algorithmic amplification

Social media has delivered the perfect amplification mechanisms for misinformation. The ability to select which individuals, groups and organizations to follow can result in echo chambers where users are exposed to a narrow sample of information and views that merely reinforce their own.

Filter bubbles provide an even greater risk for bias and polarization. Rather than users

actively choosing which sources they draw their information from, they are passively exposed to 'personalized' content determined by their social media and search engine data profiles.

"With platforms whose algorithms surface content to users outside of their social circle, individuals need not have a conspiratorial uncle or friend-of-a-friend in order to be served up misinformation in their feed – the algorithm does it for them," suggests Veena McCoolle, VP of Communications and Marketing at NewsGuard.

The Integrity Institute, helmed by former Facebook integrity team staffers Jeff Allen and Sahar Massachi, reports that content that contains [misinformation usually gets more engagement than factually accurate content.](#)

Its analysis shows, "There is typically a substantial delay between the misinformation and fact check, which is only natural since tracking down the truth takes time and effort. But this does mean that by the time the fact checks are published, the misinformation content has likely already gotten most of the engagement it ever would have."

There could be no intention to mislead, of course. Well-intentioned people might well

Misinformation has also become increasingly weaponized by political parties, state actors, and others with an agenda to push.

be unknowingly consuming and spreading misinformation without carrying out the necessary fact-checking and verification used by journalists and media-literate professionals. Certain platforms have experimented with addressing this problem in product design, like Twitter's "read before you retweet" prompt rolled out in 2020 that was aimed at curtailing misinformation virality.

Real-world problems

Medical matters, climate change and elections are examples of how false information spread online can have serious real-world implications. As Veena points out, misinformation has also become increasingly weaponized by political parties, state actors, and others with an agenda to push: "During the outbreak of COVID-19, the WHO coined the term 'infodemic' to describe the dangerous proliferation of misinformation associated with vaccines and the virus itself."

[One public health study](#) found that 52 physicians practicing in 28 different specialties across the United States propagated COVID-19 misinformation on vaccines, masks and conspiracy theories on social media and other online platforms between January 2021 and December 2022.

[The Cambridge Analytica scandal](#) serves as perhaps the most stark example of the profound impact disinformation can have on elections and democratic integrity. The firm's manipulation of data to create targeted



political advertising raised serious concerns about the influence of misleading information in swaying voter opinions and the potential to undermine the democratic process. This scandal highlighted the ways in which disinformation can be weaponized to exploit vulnerabilities in democratic systems, posing a significant threat not only to the fairness of elections but also to public trust in governance. By engineering a narrative through selective and distorted information, disinformation campaigns have demonstrated the capacity to rapidly spread, with far-reaching and sometimes irreversible consequences for society and democratic institutions.

Fake news too can have extremely damaging consequences for society and governance. Regardless of whether it's a story that contains a subtle distortion of details, only presents one viewpoint or is completely fabricated, [fake news can quickly gain traction](#).

Birdwatching: how Twitter cultivated its counter speech approach to managing misinformation

James Alexander worked on the frontline of fighting misinformation at Twitter. Find out how he and his team successfully scaled their enforcement of Twitter rules on synthetic and manipulated media and misinformation.

In the battle against misinformation, [James Alexander](#), the former Global Head of Illegal Content & Media Operations at Twitter, offers a unique frontline perspective. His experiences during the rise of misinformation in the last decade are a testament to the evolving challenges social media platforms face on an ongoing basis.

His team's early focus on synthetic and manipulated media was crucial in setting the stage for broader misinformation policies that arose later around COVID-19 and the 2020 claims of election fraud. In the beginning, James and his team believed the solution to combating misinformation would be largely automated. But as they soon found out, humans were crucial.

The 2016 US election was a turning point, as James recalls, "It's where people woke up both from an industry standpoint as well as a regulation and public standpoint." The realization that platforms could be exploited to spread misinformation that undermined something as significant as an election led to a substantial increase in resources for his team, highlighting the urgency of the issue.

James' team grappled with the delicate balance between automated and human moderation. He candidly admits, "We were entirely confident this was going to be a mostly automated method..but that was part of the biggest issue at the very beginning." The nuanced nature of misinformation required discernment that went beyond algorithms.

James' strategy was to focus on misinformation that gained visibility and could cause real-world harm. He emphasizes the importance of precision: "Knowing for certain that it is misinformation can actually be really hard... taking aim at specific known problematic misinformation is much more valuable for the resources that are required."

A significant challenge was how to balance free speech with the need to counter misinformation. In James' view, the Chinese lab leak theory around COVID-19 is a perfect example of the complexity of this task. Initially, content suggesting the virus originated from a lab in China was marked as dangerous misinformation and was suppressed. James recalls, "For a while, we did consider the lab leak as misinformation and actually as dangerous misinformation causing harm to others, so we didn't even allow it." However, as discourse and information about this theory evolved, Twitter had to adapt. "We backed off on that as more information came out...we didn't want to be tipping the scale when we didn't actually know the right answer."

In this environment of uncertainty, counter-speech emerged as a potential tool to combat misinformation without outright censorship. James suggests that sometimes the answer to misinformation may not be to silence it but to allow it to be challenged. This approach aims to provide a platform for corrective information

and dialogue, rather than removing content that is unverified or contested. It's a testament to the evolving philosophy of moderation that seeks to empower users to discern truth from falsehood, fostering a more resilient and informed online community.

James also highlights the dangerous risk of misinformation coming from influential figures given their reach and influence, stating, "I do think that the biggest risk is [misinformation coming from] somebody who has clout already."

James' approach to combating misinformation is ultimately marked by a focus on the significant majority. He cautions against the allure of the "1% of 1%" - those rare, complex cases that, while intriguing, are not representative of the broader issue. The vast majority of misinformation is more mundane and less complex, yet it has the potential to reach and influence the public on a much larger scale. "The most likely problems will always be the simplest," he notes, advocating for a moderation strategy that prioritizes the everyday experiences of the "99%." This perspective is crucial for social media platforms, in particular, directing them to invest in systems that support the vast majority of users. By concentrating on the most common and impactful forms of misinformation, platforms can allocate their resources more effectively, enhancing the overall health of the information ecosystem.

...the answer to misinformation may not be to silence it but to allow it to be challenged.

WebPurify's frontline battles on misinformation

We live in a world now where information spreads faster than ever, and this means the distinction between fact and fiction has become increasingly blurred, leading to the proliferation of misinformation. WebPurify confronts this challenge head-on, targeting misinformation that could undermine election integrity, distort public understanding of climate change, and spread falsehoods about medical issues, such as vaccine safety.

As we move into 2024 when some 40 nations will hold elections, including the United States, election-related misinformation is a significant concern. False narratives can shape voting behaviors and undermine democracy itself. WebPurify works with platforms to detect and mitigate false claims about election dates, polling locations, and voter fraud allegations. This work is crucial in safeguarding the electoral process and ensuring that the democratic fabric remains intact.

When it comes to climate change, misinformation can stall the necessary global response to an escalating crisis. Myths about climate science, deliberate understatements of human impact, and overstated claims about unproven technological solutions all serve to confuse



KING'S CHURCH INTERNATIONAL/UNSPLASH

public perception. WebPurify's role is to flag and filter out content that contradicts the scientific consensus, helping platforms stay aligned with responsible and accurate environmental reporting.

Medical misinformation has been particularly rampant in the context of the COVID-19 pandemic. Unverified home remedies, false claims about vaccine safety, and the promotion of "cures" without scientific backing not only mislead individuals but can result in direct harm to people. Social platforms continue to police COVID-19 misinformation, albeit less intensively than at the pandemic's peak, reflecting its ongoing relevance and the importance of accurate public health information.

These three examples represent the core of misinformation challenges tackled by WebPurify. The approach to each is driven by the high volume of misleading content and the severe implications of its spread. Real-world harm is the litmus test for the intensity of WebPurify's enforcement. While misinformation is an ever-evolving beast, with new topics continually emerging, the principles of protecting public discourse and welfare remain constant. The objective is clear: to curate a digital landscape where truth prevails, fostering an informed and safe online community.

"At WebPurify, we are deeply committed to ensuring the credibility and integrity

"At WebPurify, we are deeply committed to ensuring the credibility and integrity of information online"

ALEX POPKEN,
VP OF TRUST & SAFETY,
WEBPURIFY

of information online," says Alexandra Popken, WebPurify's VP of Trust & Safety. "We understand that in the digital age, the battle against misinformation is both complex and critical, which is why we use both state-of-the-art technology and expert human insight to identify and mitigate false information quickly and at scale.

"We work closely with our clients to understand their unique needs and concerns and tailor our approach in enforcing their content policies. Our goal is to create an online environment where truth prevails and communities can thrive on trust and transparency."



ADRIAN SWANCAR/UNSPASH

Future risks

In an age of misinformation, synthetic media presents perhaps the most seductive shareable content.

Synthetic media is artificially generated text, images, audio or video content that's been fully or partially created via AI algorithms. Deepfakes are one thread of synthetic media that have achieved notable mainstream awareness. The term 'deepfake' – a portmanteau of 'deep learning' and 'fake' – can be traced to 2017, when a Reddit moderator started posting videos that used face-swapping technology to add celebrity likenesses to pornographic content.



NATHAN DUMLAO/UNSPASH

Deepfakes are designed to deceive. Whether it's a digital resurrection of a deceased public figure or a former POTUS having words put in their mouth, it's possible to use the likenesses of celebrities, politicians and other prominent people to damage reputations, commit fraud and spread propaganda.

The generative AI threat

Generative AI is the creative force behind the most realistic synthetic media. It is a subset of deep learning, but one which is capable of dynamically creating new text, images, video and other content itself, based on the examples it's been shown. The results can be indistinguishable from human-generated content.

The speed and scale and at which generative AI is able to autonomously respond to unfolding events has huge implications for those charged with combating misinformation.

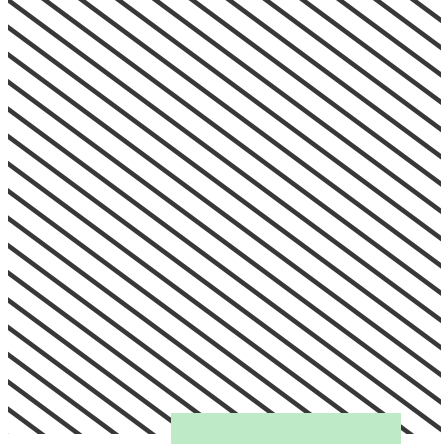
It has already presented a great threat to trust in information, Veena says: "Our research has identified an alarming propensity for generative AI chatbots to respond to prompts about topics in the news with well-written, persuasive, and entirely false accounts of the news: in some cases, complying with 100% of requests to propagate misinformation.

"In the hands of those wishing to spread harmful conspiracy theories, healthcare hoaxes, and Russian disinformation, it is a force multiplier for spreading coordinated influence operations at scale."

AI has endangered the business of journalism, Veena confirms, pointing to the alarming proliferation of Unreliable AI-generated News websites (UAINs). “We define UAINs as sites that operate with little or no human oversight, and publish articles written largely or entirely by bots,” she says.

These sites take little to no time to produce, thanks to generative AI, and many are financed entirely by programmatic advertising. UAINs are an example of ‘made for advertising’ (MFA) sites that draw advertising revenue away from deserving publishers of responsible journalism.”

AI tools for content moderation can be employed to combat these new threats. But as James Alexander, former Global Head of Illegal Content & Media Operations at Twitter underlines, they are not a magic bullet. “I think it’s really important to remember that just like with crypto or with cryptography, any weapon or backdoor can be used for both good and bad.

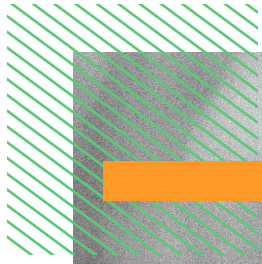


“It’s important to recognize that we’re in a constant battle against those who misuse technology to deceive.”

JAMES ALEXANDER,
FORMER GLOBAL HEAD
OF ILLEGAL CONTENT
& MEDIA OPERATIONS,
TWITTER

“The idea that technology is going to be so much better for you and will somehow not make it much easier for them to confuse you and make mistakes is a cause for concern because it’s always a bit of an arms race. That doesn’t mean not to use the technology, but it also doesn’t mean it will reduce the amount of effort and work you need to do.”

“Believing that technology will only benefit us without also making it easier for misinformation to spread is naive. It’s important to recognize that we’re in a constant battle against those who misuse technology to deceive. While embracing new technology is essential, we must also understand that it doesn’t diminish the need for vigilance and effort in combating misinformation.”



PRAVEEN THIRUMURUGAN/UNSPLASH



Strategies and Solutions

Content moderation remains a key intervention in the fight to combat misinformation. The problem is that it is, to a large extent, reactive and the effectiveness of the measures can be limited by the resources that are available.



“Too often, platforms rely on flagging instances of misinformation on a case-by-case basis,” Veena says. “This is impossible to scale, inevitably results in human error when some – but not all – content is flagged, and does not protect end users.”

By instituting policies to remove misinformation deemed inappropriate, Veena suggests that platforms can also “open themselves up to the ‘free speech’ and anti-censorship arguments that seek to preserve the rights of internet users to voice their opinions.”

Source credibility labeling can be an appropriate middle ground, she proposes, “enabling users to make their own decisions on whether a piece of content is worth sharing or trusting, based on its overarching editorial practices.”

In a significant move towards labeling to enhance transparency, Meta will soon require political advertisers to disclose if their ads were crafted using artificial intelligence. This

policy, set to be implemented in 2024 ahead of the election campaign, mandates that such ads, once vetted and approved, will be clearly labeled to indicate the use of AI tools, reflecting Meta’s commitment to combating misinformation and disinformation online.

Partnering with third parties that can provide unbiased fact-checking and/or threat intelligence around current viral misinformation campaigns is another worthwhile course of action.

“You’re not going to be an expert in everything,” James says, “so make sure that you have good partners, maybe in the news industry or in research investigations who can help lean on third-party expertise.”

Promoting digital literacy

Providing users with the critical skills they need to discern what they’re consuming is an effective proactive intervention. Veena highlights three ways that well-intentioned people can avoid being deceived by

misinformation:

- Look at the journalistic transparency and credibility of a source to make a more informed decision about whether the link or news article is something you can trust.
- Practice lateral reading and cross-reference any claims you come across in articles with trusted sources.
- Consider whether an image, video, or article you're looking at is authentic or the spawn of generative AI. Tools like GPTZero and Hive Moderation can help detect if a piece of content was AI-generated.

Prebunking – or preemptive debunking – is a technique that can be employed to help people be less susceptible to misinformation techniques.

Rather than directly debunking specific misleading claims, prebunking builds resilience by helping people to identify how they are likely to be deceived. Measures include passive methods like infographics and videos, and active methods such as online games that help people build resilience to manipulation.

Empowering online communities to counter unreliable content is another solution, particularly when faced with a high volume of viral misinformation. “I think probably the best address for scale that we’ve seen is something

like [Twitter’s] Birdwatch, or Community Notes as it is known now, which really follows the Wikipedia-type model of using crowds to collaboratively add context to potentially misleading posts,” James says.

This approach really needs a high volume of people who are vetted to make it effective, he recommends: “Birdwatch got a critical mass where it became very valuable. It is

gameable to a certain degree and there are risks to it, but you could get a lot of scale without having to focus on it and allow other people to make [it] clear when something’s not meeting muster.”

Accountability

Holding individuals and entities accountable for promoting misinformation and disinformation through effective enforcement and regulatory measures is the endgame. But there are challenges in enforcing misinformation policies when

the information available on certain topics is always changing, as Veena explains: “The COVID-19 pandemic origin is a classic example of this, when new information emerged after the fact and changed the context of previous instances of ‘misinformation.’ This is why NewsGuard is careful to only debunk provably false statements for which there is credible evidence to the contrary.”

Prebunking – or preemptive debunking – is a technique that can be employed to help people be less susceptible to misinformation techniques.



NewsGuard: fighting to improve digital literacy

In an age where misinformation spreads so rapidly, NewsGuard stands as one of the few organizations working to ensure the credibility of online news sources.

Founded in 2018, NewsGuard is primarily composed of trained journalists and information specialists who have dedicated themselves to the mission of combating the spread of false narratives online. Its approach is both meticulous and expansive; the team has collected and updated more than 6.9 million data points on more than 35,000 news and information sources. Its primary objective? To catalog and track the top false narratives proliferating on the internet.

At the heart of NewsGuard's operations is its commitment to transparency. It provides clear tools designed to counter misinformation, catering to a diverse range of users, from everyday readers to brands and even democracies. Its global team of journalists, supported by advanced AI tools, has created the trust industry's most comprehensive dataset on news. This data serves multiple purposes: it fine-tunes and establishes boundaries for generative AI models, guides brands to advertise on credible news sites while avoiding hoax platforms, offers media literacy insights for individuals, and aids democratic governments in thwarting disinformation campaigns targeting their populace.

Two of NewsGuard's standout products are its Reliability Ratings and Misinformation Fingerprints. These are designed as protective measures and fine-tuning mechanisms for AI outputs. The organization believes in the power of human judgment, especially when it comes to evaluating the credibility of news sources. Its journalist-produced data acts as a reliable guardrail, ensuring that AI models are fed trustworthy training data, thereby enhancing their accuracy.

What's more, NewsGuard has been proactive in identifying and countering the challenges posed by generative AI in the media landscape. The organization has collaborated with leading AI companies and technology platforms to ensure that their data acts as a safeguard against the potential pitfalls of AI-generated content. NewsGuard's efforts have been recognized and utilized by major players, with Microsoft, for instance, leveraging NewsGuard's trust data to enhance the reliability of its Bing Chat.

By rating the credibility of news sites and highlighting trending misinformation, NewsGuard is not only preserving the integrity of journalism but also ensuring that the public is well-informed and protected from misleading narratives.



Glossary of terms

Misinformation: Incorrect or misleading information shared without harmful intent, often as a result of honest mistakes or misunderstandings.

Disinformation: Deliberately false or misleading information spread with the intent to deceive others. Unlike misinformation, disinformation is propagated with malicious intent.

Deepfakes: Synthetic media where a person in an existing image or video is replaced with someone else's likeness using artificial neural networks. Deepfakes can make it appear as though individuals are saying or doing things they never did, which can be a form of disinformation.

Synthetic Media: Media content generated or modified using artificial intelligence (AI) technologies. This includes deepfakes, as well as other AI-generated imagery, audio, and video that can be used for benign or malicious purposes.

Fake News: A term often used to refer to fabricated news stories with false information, presented in a format mimicking traditional news outlets. It's a form of misinformation or disinformation, depending on the intent behind its creation and distribution.

Fact-Checking: The process of verifying the

accuracy and truthfulness of information, usually conducted by independent organizations or individuals dedicated to promoting factual accuracy and debunking falsehoods.

Information Disorder: A broader term encompassing various forms of distorted, misleading, or false information, including both misinformation and disinformation.

Echo Chamber: A situation in which individuals are exposed only to information from like-minded individuals, reinforcing their existing beliefs and shielding them from diverse perspectives.

Confirmation Bias: The tendency to seek, interpret, and remember information in a way that confirms one's preexisting beliefs, while giving disproportionately less consideration to alternative possibilities.

Filter Bubble: A state of intellectual isolation that can result from personalized searches when a website algorithm selectively guesses what information a user would like to see based on information about the user.

Satire: The use of humor, irony, exaggeration, or ridicule to criticize and mock people or ideas, often mistaken as true information and contributing to the spread of misinformation when taken out of context.