**KH– Civi Saclay** **Media in the digital age** **May 2024**

## Selection 8- 'Fake News' and Disinformation - AI

# Links and Resources

- **A site you might want to explore from the UNESCO Journalism: 'Fake News' and Disinformation: A Handbook for Journalism Education and Training**

*This new publication by UNESCO is a timely resource and highly topical subject for all those who practice or teach journalism in this Digital Age.*

https://webarchive.unesco.org/web/20230926213448/https://en.unesco.org/fightfakenews

Here is their full book in PDF format:

https://webarchive.unesco.org/web/20230930104950/https://en.unesco.org/sites/default/files/journalism_fake_news_disinformation_print_friendly_0.pdf

- **A few NPR podcasts**

>> China as the new national security threat (beyond Tik Tok)
https://www.npr.org/2024/04/26/1247347363/china-tiktok-national-security
>> AI and deepfakes
https://www.npr.org/2024/02/08/1229641751/ai-deepfakes-election-risks-lawmakers-tech-companies-artificial-intelligence

- **Fighting misinformation and improving media literacy**

>> https://www.npr.org/2024/03/21/1239693671/ai-images-and-conspiracy-theories-are-driving-a-push-for-media-literacy-educatio

>> See this module from the UNESCO website: **Combatting disinformation and misinformation through Media and Information Literacy (MIL)**

https://webarchive.unesco.org/web/20240306130700/https://en.unesco.org/sites/default/files/module_4.pdf

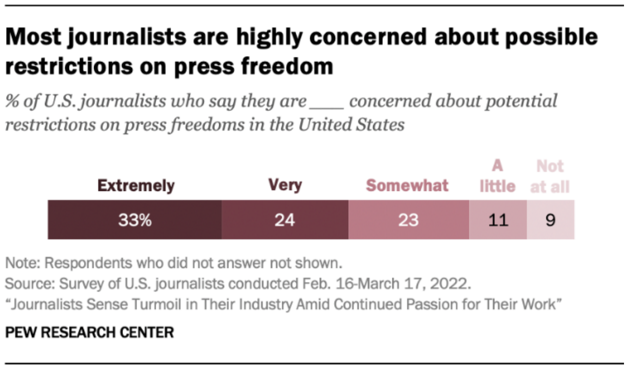
>> Here is the transcript of a very long communication **The European approach to online disinformation: geopolitical and regulatory dissonance**

**https://www.nature.com/articles/s41599-023-02179-8**

**See also next file on Legislation and Regulation**

## Document 1 - Journalists highly concerned about misinformation, future of press freedoms

Amid efforts to fight false and made-up information, anti-media campaigns, increased lawsuits and global news crackdowns, journalists in the United States express great concern about the future of press freedoms.



**Most journalists are highly concerned about possible restrictions on press freedom**

*% of U.S. journalists who say they are ____ concerned about potential restrictions on press freedoms in the United States*

Note: Respondents who did not answer not shown.
Source: Survey of U.S. journalists conducted Feb. 16-March 17, 2022.
"Journalists Sense Turmoil in Their Industry Amid Continued Passion for Their Work"

PEW RESEARCH CENTER

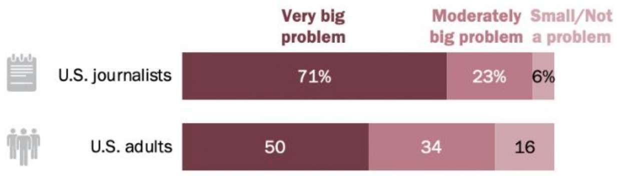
Roughly six-in-ten journalists surveyed say they are either extremely (33%) or very concerned (24%) about potential restrictions on press freedoms in the U.S.

About a quarter (23%) are somewhat concerned, while just one-in-five express low levels of concern about this.

Journalists see false and made-up news as a big problem and don't have much confidence in how the industry handles it

**Journalists much more likely than the American public to see made-up news as a major concern**

*% who say that made-up news and information is a ____ in the country today*



Note: Respondents who did not answer not shown.
Source: Survey of U.S. journalists conducted Feb. 16-March 17, 2022. Survey of U.S. adults conducted March 7-13, 2022.
"Journalists Sense Turmoil in Their Industry Amid Continued Passion for Their Work"
PEW RESEARCH CENTER

Another area of concern for journalists is the volume of erroneous information today. More than nine-in-ten journalists surveyed (94%) say made-up news and information is a significant problem in America today, with 71% identifying it as a very big problem and 23% seeing it as a moderately big problem; 6% say it is a small problem or not a problem at all.

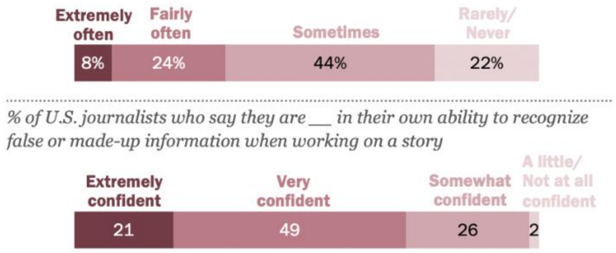
The American public also sees made-up news and information as a problem, but not quite to the same extent. In a separate survey of 10,441 U.S. adults conducted March 7-13, 2022, 50% say made-up news is a very big problem (21 percentage points below journalists), while another 34% say it is a moderately big problem and 16% say it is a small problem or not a problem at all.

Misinformation is a fairly regular topic of conversation within the newsroom itself. About six-in-ten journalists (58%) say they had conversations with colleagues about misinformation at least several times a month over the past year.

**Most journalists come across made-up information when working on story and are confident they can recognize it**

*% of U.S. journalists who say they __ come across information they believe is false or made up when working on a story*

*% of U.S. journalists who say they are __ in their own ability to recognize false or made-up information when working on a story*



Note: Respondents who did not answer not shown.
Source: Survey of U.S. journalists conducted Feb. 16-March 17, 2022.
"Journalists Sense Turmoil in Their Industry Amid Continued Passion for Their Work"
PEW RESEARCH CENTER

The survey also finds that one-third of journalists indicate that they deal with false or made-up news in their work on a fairly regular basis – saying that they come across false information when working on a story either extremely often (8%) or fairly often (24%). Another 44% say they sometimes come across it.

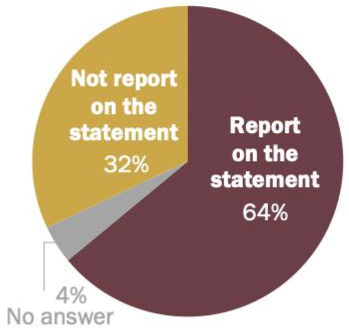
About seven-in-ten journalists (71%) say they are either extremely (21%) or very confident (49%) in their ability to recognize false information when they are working on a story.

Still, specifically among reporting journalists, about a quarter (26%) say they unknowingly reported on a story that was later found to contain false information. (Reporting journalists are those who indicated in the survey that they report, edit or create original news stories *and* that they have one of the following job titles: reporter, columnist, writer, correspondent, photojournalist, video journalist, data visualization journalist, host, anchor, commentator or blogger. About three-quarters of all journalists in this study – 76% – are reporting journalists.)

While the journalists surveyed here may feel good about their own ability to detect misinformation, they are not particularly confident in the industry's ability to manage or correct it. Only 8% of all journalists surveyed say news organizations do a very good job at handling misinformation, while another 35% say news outlets are somewhat good at it – lower than the ratings journalists give news organizations on several other core functions (see Chapter 4).

And most say their news organization (or the main one they work for if they work for more than one) does *not* have formal guidance on how to handle made-up and false information in their jobs. Six-in-ten say their organization does not have guidelines for how to handle false and made-up information that they come across, far higher than the 36% who say their organization does.

Most journalists think it is important to report on the false statements of public figures



**Most journalists say news organizations should report on public figures' false statements**

*% of U.S. journalists who say that if a public figure makes a statement that is false or made up, news organizations should ...*

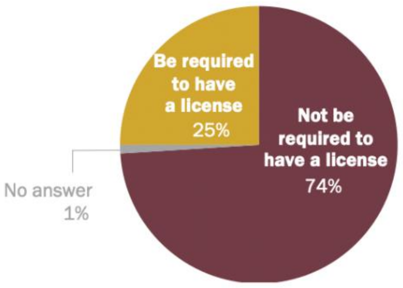
Source: Survey of U.S. journalists conducted Feb. 16-March 17, 2022.
"Journalists Sense Turmoil in Their Industry Amid Continued Passion for Their Work"

**PEW RESEARCH CENTER**

Most journalists think that part of managing misinformation means reporting on public figures who make false or made-up statements. Twice as many journalists say that if a public figure makes a statement that is false or made up, news organizations should "report on the statement because it is important for the public to know about" (64%) rather than "not report on the statement because it gives attention to the falsehoods and the public figure" (32%).

Many journalists, though, never or almost never cover the issue of misinformation. Two-thirds of journalists surveyed (66%) say almost none of the stories they worked on in the past year had to do with false or made-up information. Just 6% of those surveyed say half or more of the news stories they worked on covered false or made-up news in some way, while about a quarter (27%) say that some of their stories – but fewer than half – dealt with this topic.

Vast majority of journalists are against requiring a license to be a journalist



**About three-quarters of journalists are against requiring licenses to practice journalism**

*% of U.S. journalists who say that people should ___ to practice journalism*

Source: Survey of U.S. journalists conducted Feb. 16-March 17, 2022.
"Journalists Sense Turmoil in Their Industry Amid Continued Passion for Their Work"

**PEW RESEARCH CENTER**

One particular feature of journalism is that there is no requirement to have a license or certification process to call oneself a journalist – unlike a physician would in order to practice medicine in the United States. The question of whether to require one or not occasionally gets raised. As of now, a solid majority of journalists are against such a requirement. Nearly three-quarters of journalists (74%) are in favor of continuing to allow journalists to practice journalism without needing a license. One-quarter of U.S. journalists would like to see a license required for members of their industry. (Chapter 8 looks at how these views vary by the original platform of the organization that journalists work for.)

Currently, there is no licensing requirement for journalists themselves. Radio and television stations are licensed and regulated by the Federal Communications Commission, but there is no such regulatory authority for newspapers and online outlets.

# Document 2 - **How disinformation works—and how to counter it**

*More co-ordination is needed, and better access to data*

Leader, ***The Economist,*** May 4th 2024

Did you know that the wildfires which ravaged Hawaii last summer were started by a secret “weather weapon” being tested by America’s armed forces, and that American ngos were spreading dengue fever in Africa? That Olena Zelenska, Ukraine’s first lady, went on a $1.1m shopping spree on Manhattan’s Fifth Avenue? Or that Narendra Modi, India’s prime minister, has been endorsed in a new song by Mahendra Kapoor, an Indian singer who died in 2008?

These stories are, of course, all bogus. They are examples of disinformation: falsehoods that are intended to deceive. Such tall tales are being spread around the world by increasingly sophisticated campaigns. Whizzy artificial-intelligence (AI) tools and intricate networks of social-media accounts are being used to make and share eerily convincing photos, video and audio, confusing fact with fiction. In a year when half the world is holding elections, this is fuelling fears that technology will make disinformation impossible to fight, fatally undermining democracy. How worried should you be?

Disinformation has existed for as long as there have been two sides to an argument. Rameses II did not win the battle of Kadesh in 1274bc. It was, at best, a draw; but you would never guess that from the monuments the pharaoh built in honour of his triumph. Julius Caesar’s account of the Gallic wars is as much political propaganda as historical narrative. The age of print was no better. During the English civil war of the 1640s, press controls collapsed, prompting much concern about “scurrilous and fictitious pamphlets”.

The internet has made the problem much worse. False information can be distributed at low cost on social media; ai also makes it cheap to produce. Much about disinformation is murky. But in a special Science & technology section, we trace the complex ways in which it is seeded and spread via networks of social-media accounts and websites. Russia’s campaign against Ms Zelenska, for instance, began as a video on YouTube, before passing through African fake-news websites and being boosted by other sites and social-media accounts. The result is a deceptive veneer of plausibility.

Spreader accounts build a following by posting about football or the British royal family, gaining trust before mixing in disinformation. Much of the research on disinformation tends to focus on a specific topic on a particular platform in a single language. But it turns out that most campaigns work in similar ways. The techniques used by Chinese disinformation operations to bad-mouth South Korean firms in the Middle East, for instance, look remarkably like those used in Russian-led efforts to spread untruths around Europe.

The goal of many operations is not necessarily to make you support one political party over another. Sometimes the aim is simply to pollute the public sphere, or sow distrust in media, governments, and the very idea that truth is knowable. Hence the Chinese fables about weather weapons in Hawaii, or Russia’s bid to conceal its role in shooting down a Malaysian airliner by promoting several competing narratives.

All this prompts concerns that technology, by making disinformation unbeatable, will threaten democracy itself. But there are ways to minimise and manage the problem.

Encouragingly, technology is as much a force for good as it is for evil. Although AI makes the production of disinformation much cheaper, it can also help with tracking and detection. Even as campaigns become more sophisticated, with each spreader account varying its language just enough to be plausible, AI models can detect narratives that seem similar. Other tools can spot dodgy videos by identifying faked audio, or by looking for signs of real heartbeats, as revealed by subtle variations in the skin colour of people's foreheads.

Better co-ordination can help, too. In some ways the situation is analogous to climate science in the 1980s, when meteorologists, oceanographers and earth scientists could tell something was happening, but could each see only part of the picture. Only when they were brought together did the full extent of climate change become clear. Similarly, academic researchers, NGOs, tech firms, media outlets and government agencies cannot tackle the problem of disinformation on their own. With co-ordination, they can share information and spot patterns, enabling tech firms to label, muzzle or remove deceptive content. For instance, Facebook's parent, Meta, shut down a disinformation operation in Ukraine in late 2023 after receiving a tip-off from Google.

But deeper understanding also requires better access to data. In today's world of algorithmic feeds, only tech companies can tell who is reading what. Under American law these firms are not obliged to share data with researchers. But Europe's new Digital Services Act mandates data-sharing, and could be a template for other countries. Companies worried about sharing secret information could let researchers send in programs to be run, rather than sending out data for analysis.

Such co-ordination will be easier to pull off in some places than others. Taiwan, for instance, is considered the gold standard for dealing with disinformation campaigns. It helps that the country is small, trust in the government is high and the threat from a hostile foreign power is clear. Other countries have fewer resources and weaker trust in institutions. In America, alas, polarised politics means that co-ordinated attempts to combat disinformation have been depicted as evidence of a vast left-wing conspiracy to silence right-wing voices online.

**One person's fact...**

The dangers of disinformation need to be taken seriously and studied closely. But bear in mind that they are still uncertain. So far there is little evidence that disinformation alone can sway the outcome of an election. For centuries there have been people who have peddled false information, and people who have wanted to believe them. Yet societies have usually found ways to cope. Disinformation may be taking on a new, more sophisticated shape today. But it has not yet revealed itself as an unprecedented and unassailable threat. ■

**See the related article at the end of the longer online version of this file**

## Document 3 - **The vocabulary of disinformation**

The Economist explains, May 2nd 2024

The words disinformation and misinformation are often used interchangeably. But there is a subtle difference in their meaning. Disinformation is when false information is spread with the intent to deceive, often as part of a co-ordinated campaign. Misinformation is when false information is unintentionally spread, for instance by repeating a rumour to a friend or reposting an unsubstantiated claim on a social-media platform.

Disinformation is not a new phenomenon, but social media and artificial intelligence (ai) are making it easier to spread lies. A disinformation campaign might try to sway voters in the run-up to an election, turn social groups against one another, undermine scientific research, discredit a business or manipulate share prices. As technology makes the spread of false information ever more complex, disinformation hunters are fighting back—and the language used to talk about disinformation is changing.

**ai-generated news**

Misleading articles produced by ai for made-up-news sites or content farms. At first glance the websites that host such stories may look similar to those of real media outlets. They often run innocuous stories about travel or entertainment alongside articles that peddle harmful falsehoods. See how these sites work in practice here. Not to be confused with attempts by some ai firms to train their large-language models to write legitimate stories about the news.

**Alt-tech**

A broad term used to refer to websites, including social-media platforms, that are not mainstream. Many such social-media platforms, including Gab and Parler, have found favour among fringe groups because they have loose rules about the type of content allowed on their sites. As a result they can be hotbeds of false information.

**Bot**

A social-media account programmed to perform a certain action. Bots (originally short for "robot") can be useful, and even fun. One bot on x gathers linked posts together and presents them in an easily readable format; another on Reddit reposts messages as haikus. But they are often used for nefarious purposes: a bot may be set up to appear like the account of an ordinary person, then harass other users, amplify falsehoods or trick others into clicking on scam links. Social-media platforms try to detect and minimise their activity, but many bots slip through the cracks.

**captcha test**
An initialism for "Completely Automated Public Turing test to tell Computers and Humans Apart". captchas are used by various websites to try to identify bots. They ask users to perform a task that a computer is supposed to be unable to do, such as to type out a sequence of letters, or to identify a set of matching visual cues. But as technology improves, bots are starting to outsmart them.

**Catfish**
A person who creates a fake social-media profile, posing as someone else. Some create a range of fake profiles to support their made-up identity with fictional friends and family. Catfish typically try to form an emotional connection with other social-media users for the purpose of harassment of financial gain. Catfish are sometimes also referred to as "sock puppet" accounts.

**Co-ordinated inauthentic behaviour (CIB)**
Using software to control hundreds or thousands of social-media accounts, usually to spread disinformation. Such accounts may all send out the same message, or like and share a given post. CIB manipulates social-media algorithms by suggesting there is widespread interest in a particular viewpoint: this means a site is more likely to show posts that share that viewpoint to other users.

**Content farm**
Websites that rely on low-paid writers or ai to churn out articles, with the goal of becoming highly ranked by search engines in order to boost revenue from advertisers. They may also be used to spread false information.

**Content moderation**
Most social-media platforms set rules for good behaviour on their sites, known as "community standards". If a user breaks a given rule—by making hateful comments, for example, or spreading lies—they may be punished with a temporary suspension or permanent ban. Determining what does or does not count as breaking the rules falls under the remit of content moderators employed by these sites. Their work is becoming increasingly contentious.

**Corrective information**
Using fact-checkers to debunk disinformation. Many social-media platforms have in-house teams dedicated to fact-checking posts that are flagged to them by users as potentially false. If they are found to contain deceptive content, they may be appended with a label explaining what they have got wrong, or taken down.

**Dark web**
A hidden, encrypted layer of the internet that is not discoverable on mainstream search engines, instead requiring special browsers to reach it. The dark web is not illegal to access—in fact it was created as a way to share information and files freely online, and has been used by America's military—but illegal activity is known to take place there. Some dark-web sites offer services such as hacking, identity theft and the creation of "deepfakes".

**Deepfake**
An ai-generated image or video that convincingly shows a person doing or saying something they have not, by superimposing their face onto the body of someone else or generating an entirely new visual. As ai tools to parrot people's voices have been developed, the malicious use of fake audio is sometimes also described as a deepfake. The term is named after a Reddit account which, in 2017, shared fake pornographic videos of female celebrities. Some fake images and videos are made without the use of ai, for instance by using social-media "filters" (pre-made effects that distort an image) or photo-editing software. Such efforts are sometimes termed "cheapfakes".

**Doxxing**
Spreading true information with an intention to cause harm, usually by sharing personal information about a target, such as their address, online. Such leaks are often intended to fuel harassment.

**Fake news**
This term was once used to describe disinformation. But it has become politicised in recent years, in large part because Donald Trump has frequently used the label to attack his legitimate critics. It is now frequently used as a retort to dismiss verified facts as untruths.

**Infodemic**
Defined by the World Health Organisation as a crisis of "too much information", both real and false. This can be especially harmful during a rapid outbreak of disease, such as the covid-19 pandemic, or when wars begin, because it becomes hard to determine what is and is not true. Faced with an overwhelming number of posts on social media, people may focus on whatever they see most prominently or what their friends or family share, even if the information is untrue.

**Integration**
When credible sources and genuine accounts pick up disinformation and share it, either inadvertently or on purpose.

**Impostor account**
A social-media account used to infiltrate a political movement online. Creators of these accounts start by pretending to be on the side of protestors to build up a following—then post disinformation about the cause to discredit it.

**Layering**
A process in a disinformation campaign that involves creating a trail from the original source of false information to a more credible one. The more layers involved, the stronger the appearance of the same narrative from multiple sources—and the more convincing the lie.

**Malinformation**
The spread of verified information with the intent to cause harm or manipulate. This might be done by leaking personal information about someone to endanger them (see "doxxing"), or by sharing something true but removing some of its context—by cutting frames out of a video or cropping a photo, for instance—to confuse recipients.

**Meme**
Visual creations, ideas or inside jokes that are spread rapidly and replicated among those with similar interests, often without identifying the original creator. Because they can resonate with a target audience and go viral they have become a popular tool for spreading false information and propaganda.

**Microtargeting**
Web users' online habits are tracked, painting a virtual portrait of what sort of person they might be—and what they might buy. These data can be bought by advertisers to put personalised messages in front of certain groups. But they can also be used to sharpen a disinformation campaign by targeting specific users based on their preferences and biases. ai can bolster microtargeting by hyper-personalising the type of content a user might see.

**Placement**
The initial posting of a fabricated piece of media or lie online, typically through an anonymous or false account, to kickstart a disinformation campaign. See also "seeders".

**Pre-bunking**
Using media-literacy education as an inoculation method against disinformation. The idea is that fact-checkers identify a false narrative that is starting to circulate and make people aware of it early, so that if they do encounter it elsewhere on social media, they will do so with scepticism.

**Seeders**
Social-media accounts or sites that plant disinformation, for instance by posting or sharing links to a dodgy article. On social media, these accounts typically have just a small number of followers. . Because of their size, they have little influence in disseminating the false information. Instead, they rely on spreader accounts, which have a larger following, to amplify the stories.

**Sock puppet**
See "catfish".

**Spamouflague**
The name of a Chinese state-backed propaganda group, which also goes by Dragonbridge and Storm-1376, that disseminates disinformation. The word also refers to a tactic used by spreader accounts (defined below) to disguise disinformation by censoring words or sprinkling in harmful posts among innocuous ones to avoid detection by tech platforms and to appear legitimate.

**Spreader**
Spreader social-media accounts have large followings and are used to reshare posts planted on seeder accounts, to amplify disinformation. They typically collect followers and avoid detection by mixing in posts about popular, unrelated topics, such as football, or by sharing images of scantily-clad women. In many poor countries there is a growing cottage industry of cultivating spreader accounts and selling them on to bad actors.

**Troll**
A deliberately antagonistic social-media user who posts upsetting or provocative content online to elicit a reaction from others. Most trolls do not operate under their real names. Some volunteers are attempting to fight back: in 2014 a group of fact-checkers in Lithuania set out to counteract pro-Kremlin trolls. They named themselves "elves".

**Troll factory**
An organised group of trolls, hired to create havoc or interfere with political discourse online. In 2018 America imposed sanctions on the Internet Research Agency, a Russian troll factory founded by Yevgeniy Prigozhin, the late boss of the Wagner Group, for election interference.

**Verification**
Processes used to certify the authenticity of social-media accounts belonging to news outlets, businesses, politicians and celebrities, in an attempt to prevent impersonators from spreading false information under the names of these people or brands. Some countries are considering passing laws that would require users to submit official identification, such as a passport or a driving licence, to social-media platforms to confirm their identity—but that could make things harder for users living under oppressive governments. ■

## Document 4 - HOW JOURNALISTS CAN COMBAT POLITICAL DISINFORMATION IN A WORLD OF ECHO CHAMBERS AND DEEPFAKES

Pen America, May 7, 2024, *By Mina Haq*

Journalists face a daunting task ahead of the 2024 election as disinformation campaigns grow more sophisticated and public trust in institutions – including the news media – declines. A new Associated Press-American Press Institute poll revealed 53% of Americans say they are extremely or very concerned that news organizations will report inaccuracies or misinformation during the election.

The National Press Club Journalism Institute hosted a conversation on May 1 between journalists and experts who laid out the scope of the disinformation problem – and how reporters can combat it. It was the fourth and final webinar in a training series focused on ethics and disinformation.

As the election season ramps up, panelists emphasized the importance of building relationships with local election officials and prioritizing explanatory journalism that speaks to readers' concerns, especially as micro-targeted disinformation campaigns seek out vulnerable communities.

Votebeat editor-in-chief Chad Lorenz moderated the discussion between Sheera Frenkel, a New York Times technology reporter; Christine Fernando, an Associated Press democracy reporter; Yaël Eisenstat, a senior fellow at Cybersecurity for Democracy and a PEN America consultant; and Tina Barton, a senior elections expert for the Committee for Safe and Secure Elections.

### Pervasive falsehoods and how they spread

Election mis- and disinformation isn't new. False narratives about stolen elections, rigged voting machines and tossed ballots have spread for years, thanks to a combination of bad actors purposely spreading false information and unwitting participants repeating what they believe to be true. But there are also new problems created by a decentralized social media environment and the rapid rise of generative AI.

"People are getting their information from lots of different places, and that creates a challenge for reporters," Frenkel said.

News consumers aren't just on Facebook and Twitter anymore. Frenkel cited far-right social networks like Gab and Parlor as places where election-related conspiracy theories can originate, then spread across platforms, creating a challenge for reporters trying to track and get ahead of false narratives.

This scattered internet environment, where like-minded communities often gather in one space, also creates openings for disinformation that targets vulnerable communities about historically relevant and emotionally resonant subjects, such as authoritarianism or communism. Disinformers "prey on the very specific traumas and fears of certain communities," Fernando said, and can take advantage of language barriers and information gaps.

The rise of generative AI also creates a new problem for reporters. News organizations must add debunking fake audio and video – which they did in the case of the President Biden deepfake robocall targeting New Hampshire voters – to their list of duties, but it could also be another blow to the public's trust. Forty-two percent of Americans expressed worry in the AP poll that news outlets will use generative AI to create stories.

It's not that journalists aren't capable of debunking AI-generated content, Eisenstat said. But newsrooms will have to explain to readers what these tools do and how they're being used to mislead. That becomes difficult when audiences are primed to believe the narrative – real or not – that confirms preexisting beliefs.

"The awareness that these tools exist actually causes more distrust in information to begin with, and that is going to be really concerning and a really big thing for journalists to tackle," Eisenstat said.

### Practical tips to fight disinformation

How can journalists and elections officials tackle these issues? Transparency and immersion within specific communities are key.

Barton encouraged journalists to build relationships with local elections officials and immerse themselves in the nuances of the process so they can accurately and clearly relay information to readers.

“They need you in those times to be a voice for them,” she said to journalists.

Frenkel suggested journalists embed themselves into communities where dis- and mis-information originate, allowing them to determine which narratives are circulating and what might necessitate “prebunking.” To prebunk is to “preemptively refute expected false narratives, misinformation or manipulation techniques,” according to the Poynter Institute.

Other panelists echoed the importance of prebunking: “It’s priming your audience to already have information at their fingertips when those narratives hit,” Eisenstat said.

Panelists urged caution to avoid platforming misinformation while covering it. At the New York Times, Frenkel said, editors and reporters frequently debate whether to write about the latest false narrative. She said it meets the coverage threshold when it has reached a critical mass of people or is being shared by public figures, but that reporters should lead with the fact that the information is false. To build trust, it also helps to explain vetting processes, the organization’s sourcing standards, and other work that goes into reporting a story.

“I can’t count the number of people that seem shocked by the amount of work we do to get one verified claim,” Frenkel said.

Adding context is also crucial when covering false information, Eisenstat said, keeping in mind how purveyors of disinformation can screenshot or use accurate reporting to perpetuate their own agenda.

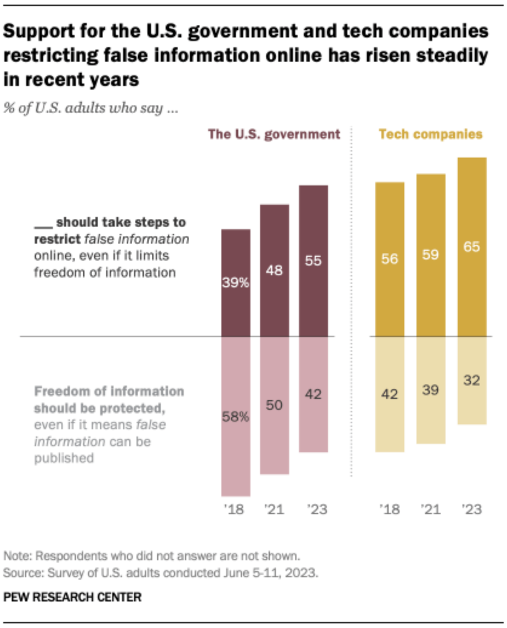
It would be impossible – and irresponsible – for reporters to cover every instance of mis- or disinformation, but there are other ways to mitigate its spread. Fernando said she has noticed several grassroots efforts by organizations working with local election officials to debunk misinformation by meeting people where they are, such as through community events or with Spanish-language media. But, she added, it’s important to understand that it takes time to reach people.

“An election-denier isn’t going to read one story and be like, ‘Okay, I believe you.’ It takes a lot of time and a lot of trust.”

## Document 5 -**Most Americans favor restrictions on false information, violent content online**

**PEW REARCH CENTER, JULY 20, 2023**

Most Americans say the U.S. government and technology companies should each take steps to restrict false information and extremely violent content online. However, there is more support for tech companies moderating these types of content than for the federal government doing so, according to a new Pew Research Center survey.


**Support for the U.S. government and tech companies restricting false information online has risen steadily in recent years**

*% of U.S. adults who say ...*

Note: Respondents who did not answer are not shown.
Source: Survey of U.S. adults conducted June 5-11, 2023.
PEW RESEARCH CENTER

Support for both technology companies and the government taking steps to restrict false information online has grown in recent years. For example, the share of U.S. adults who say the federal government should restrict false information has risen from 39% in 2018 to 55% in 2023.

This increase in support comes amid public debates about online content regulation and court cases that look at how tech companies moderate content on their platforms.

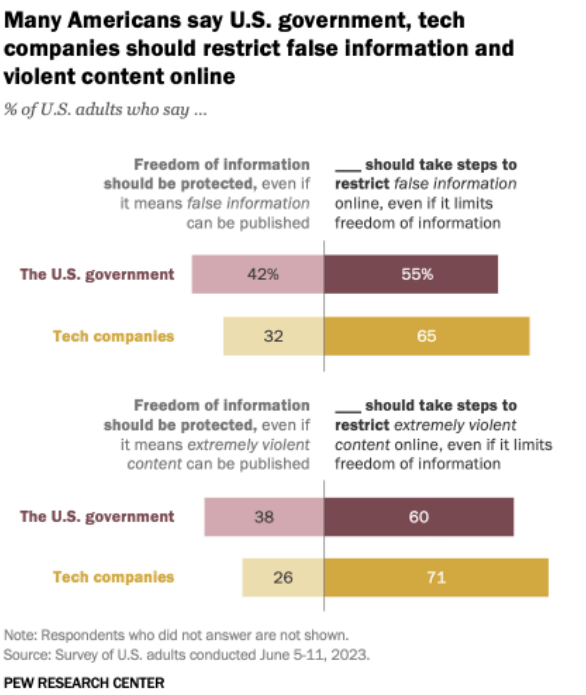
Additionally, tech companies have begun to remove some content restrictions that they had imposed in response to misinformation about the COVID-19 pandemic and the 2020 election.

That said, the amount that people have heard about the debates surrounding the role government should play in regulating major technology companies has decreased in the past two years. In 2021, 51% of U.S. adults said they had heard at least a fair amount about this topic, compared with 39% today.

### *Key takeaways*

- 65% of Americans support tech companies moderating false information online and 55% support the U.S. government taking these steps. These shares have increased since 2018.
- Americans are even more supportive of tech companies (71%) and the U.S. government (60%) restricting extremely violent content online.
- Democrats are more supportive than Republicans of tech companies and the U.S. government restricting extremely violent content and false information online. The partisan gap in support for restricting false information has grown substantially since 2018.

*Views toward moderating false information online*



Many Americans say U.S. government, tech companies should restrict false information and violent content online

*% of U.S. adults who say ...*

Note: Respondents who did not answer are not shown.
Source: Survey of U.S. adults conducted June 5-11, 2023.
PEW RESEARCH CENTER

Just over half of Americans (55%) support the **U.S. government taking steps to restrict false information online,** even if it limits people from freely publishing or accessing information.

U.S. adults are less likely to say that freedom of information should be protected even if it means false information can be published (42%).

Support for government intervention has steadily risen since the first time we asked this question in 2018. In fact, the balance of opinion has tilted: Five years ago, Americans were more inclined to prioritize freedom of information over restricting false information (58% vs. 39%).

In addition, the share of U.S. adults who say that **tech companies should take steps to restrict false information online** has increased from 56% in 2018 to 65% in 2023.

### *Attitudes on moderating extremely violent content online*

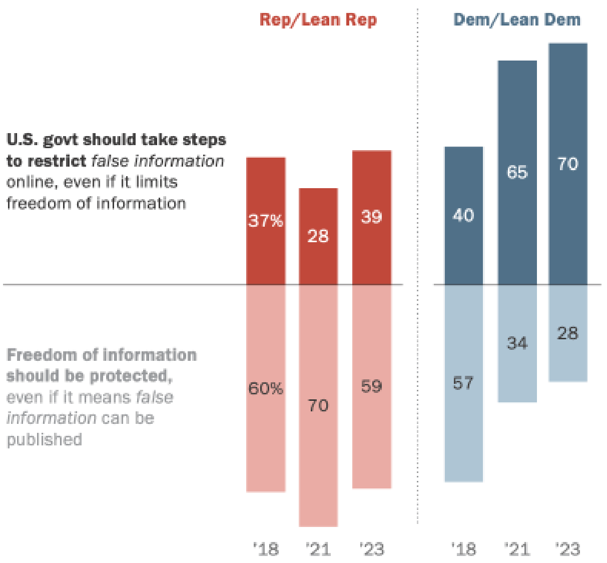
This was the first time we asked about the American public’s views of moderating extremely violent content online. We found that Americans are somewhat more likely to favor restricting this type of content than false information.

About seven-in-ten Americans (71%) believe that tech companies should restrict violent content online, and 60% say that the government should do so.

*Partisan differences*

**Democrats have become much more supportive than Republicans of the government restricting false information online**

*% of U.S. adults who say …*



Note: Respondents who did not answer are not shown.
Source: Survey of U.S. adults conducted June 5-11, 2023.
PEW RESEARCH CENTER

Democrats and Democratic-leaning independents are much more likely than Republicans and Republican leaners to support the U.S. government taking steps to **restrict false information online** (70% vs. 39%).

There was virtually no difference between the parties in 2018, but the share of Democrats who support government intervention has grown from 40% in 2018 to 70% in 2023, while the share of Republicans who hold this view hasn't changed much.

There is a similar gap between the shares of Democrats and Republicans who say technology companies should restrict false information online.

A large majority of Democrats and Democratic leaners (81%) support technology companies taking such steps, while about half of Republicans (48%) say the same. The share of Democrats who support technology companies taking these steps has also increased steadily since 2018.

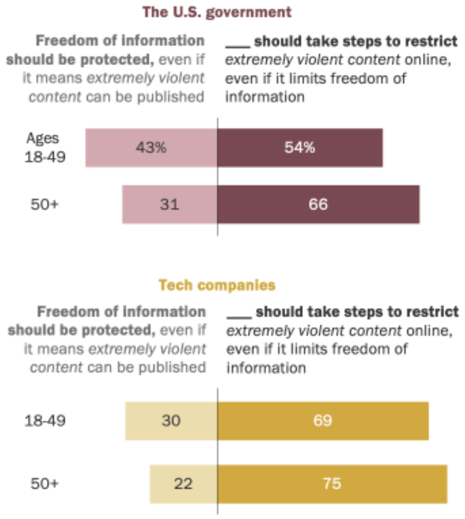
These partisan gaps persist when it comes to restricting **extremely violent content online**.

Democrats are more likely than Republicans to say that the U.S. government (71% vs. 48%, respectively) and tech companies (83% vs. 61%) should take steps to restrict violent content online even if it limits freedom of information.

*Age differences*

**Older Americans more likely than younger Americans to support restricting extremely violent content online**

*% of U.S. adults who say …*



Note: Respondents who did not answer are not shown.
Source: Survey of U.S. adults conducted June 5-11, 2023.
PEW RESEARCH CENTER

U.S. adults ages 50 and older are more likely than younger adults to say that both technology companies (68% vs. 62%) and the U.S. government (58% vs. 52%) should take steps to **restrict false information online**.

However, the shares of younger adults who say they support tech companies and the government restricting false information online have increased substantially since 2018 (by 14 and 19 percentage points, respectively).

There are similar divides when it comes to **restricting violent content online**.

Three-quarters of Americans ages 50 and older support tech companies restricting violent content online, and 66% support the U.S. government doing so. This compares with 69% and 54%, respectively, among younger adults.

## Document 5 - **Inside Google's Plans to Combat Misinformation Ahead of the E.U. Elections**

BY YASMEEN SERHAN, ***TIME***, APRIL 25, 2024

"We are all at risk of manipulation online right now."

So begins a short animated video about a practice known as decontextualization and how it can be used to misinform people online. The video identifies signs to watch out for, including surprising or out of the ordinary content, seemingly unreliable sources, or video or audio that appear to have been manipulated or repurposed.

Though it may not look like it, this 50-second video is actually an election ad—one of three that Google will be rolling out across five European countries next month in advance of the European Union's June parliamentary elections. But unlike traditional election ads that are designed to persuade people how to vote, these are seeking to educate voters about how they could be misled. It's an initiative that Google describes as preventative debunking—or, more simply, "prebunking."

"It works like a vaccine," Beth Goldberg, the head of research at Google's internal Jigsaw unit, which was founded in 2010 with a remit to address threats to open societies, tells TIME. By enabling prospective voters to recognize common manipulation techniques that could be used to mislead them—such as scapegoating or polarization—Goldberg says that prebunking "helps people to gain mental defenses proactively."

Concerns about AI-generated disinformation and the impact it stands to have on contests around the world continues to dominate this year's election megacycle. This is particularly true in the E.U., which recently passed a new law compelling tech firms to increase their efforts to clamp down on disinformation amid concerns that an uptick in Russian propaganda could distort the results.

Contrary to what one might expect, prebunking ads aren't overtly political nor do they make any allusions to any specific candidates or parties. In the video about decontextualization, for example, viewers are shown a hypothetical scenario in which an AI-generated video of a lion set loose on a town square is used to stoke fear and panic. In another video, this time about scapegoating, they are shown an incident in which a community lays sole blame on another group (in this case, tourists) for the litter in their parks without exploring other possible causes.

The beauty of this approach, Goldberg notes, is that it needn't be specific. "It doesn't have to be actual misinformation; you can just show someone how the manipulation works," she says, noting that keeping the content general and focusing on manipulation strategies, rather than the misinformation itself, allows these campaigns to reach people regardless of their political persuasion.

While Google's prebunking campaign is relatively new, the tactic is not. Indeed, the concept dates back to the 1960s, when the social psychologist William McGuire sought to understand people's susceptibility to propaganda during the Cold War and whether they could be defended against it. This culminated in what McGuire called "inoculation theory," which rested on the premise that false narratives, like viruses, can be contagious and that by inoculating people with a dose of facts, they can become less susceptible. But it wasn't until decades later that the theory began being applied to online information. In recent years, Jigsaw has conducted prebunking initiatives in Eastern Europe and Indonesia. Its forthcoming European campaign, which formally kicks off in May, will primarily be disseminated as short ads on YouTube and Meta platforms targeting voters in Belgium, France, Germany, Italy, and Poland. Afterwards, viewers will be invited to take a short, multiple-choice survey testing their ability to identify the manipulation technique featured in the ad.

Whereas prebunking doesn't necessarily face as much resistance as more conventional forms of combating misinformation such as fact checking or content moderation, which some critics have likened to censorship, it isn't a panacea either. Jon Roozenbeek, an assistant professor in psychology and security at King's College London who has spent years working with Jigsaw on prebunking, tells TIME that one of the biggest challenges in these campaigns is ensuring that the videos are captivating enough to hold viewers' attention. Even if they do, he adds, "You can't really expect miracles in a sense that, all of a sudden after one of these videos, people begin to behave completely differently online." he says. "It's just way too much to expect from a psychological intervention that is as light touch as this."

This isn't to say that prebunking doesn't have an impact. In previous campaigns, post-ad surveys showed that the share of individuals who could correctly identify a manipulation technique increased by as much as 5% after viewing a prebunking video. “We’re not doubtful that the effect is real; it’s just you can argue over whether it’s large enough,” Roozenbeek says. “That’s the main discussion that we’re having.”

While Jigsaw has led the way on prebunking efforts, they’re not the only ones utilizing this approach. In the U.S., the Biden administration has sought to counter Russian disinformation in part by declassifying intelligence forecasting the kinds of narratives that it anticipated the Kremlin would use, particularly in the run up to Moscow’s 2022 full-scale invasion of Ukraine. This practice has since extended to China (where the U.S. government used declassified materials to forecast potential Chinese provocations in the Taiwan Strait) and Iran (the U.S. declassified intelligence claiming that Tehran had transferred drones and cruise missiles to Houthi militants in Yemen that were being used to attack ships in the Red Sea). What the White House has billed as strategic declassification is just prebunking by another name.

Working with academics and civil society organizations across the E.U.’s 27 member states, Jigsaw’s latest prebunking campaign is set to be its biggest and most collaborative effort yet. And in an election that will see hundreds of millions of voters go to the polls to elect what polls project could be the most far-right European Parliament today, the stakes couldn’t be higher.

## Document 6 - **How ‘fighting disinformation’ turns into political censorship**

*Self-appointed monitors can financially harm publications as they choose.*

By Freddie Sayers, , May 8, 2024

*Freddie Sayers is the editor in chief of UnHerd.*

UnHerd, the Britain-based publication I lead, published an investigation on April 17 into a transatlantic organization called the Global Disinformation Index. We revealed that, having received money from the U.S. State Department, as well as the British, German and European Union governments, the GDI issues what amount to blacklists of news publications, on highly tendentious grounds, that online advertising exchanges then consult and can use to justify turning off ad revenue.

With worries about the rise of “disinformation” in recent years, various projects were launched in the United States, Britain and elsewhere, many no doubt with good intentions, to combat disinformation’s deleterious effects on democratic values. What has emerged, though, is an opaque network of private and government-supported enterprises that appear intent on censoring political views they find unpalatable.

Just last month saw the U.S. launch of a new effort billing itself as combating disinformation — but its political agenda was unmistakable. The American Sunlight Project is the brainchild of Nina Jankowicz. She headed President Biden’s Disinformation Governance Board for the three weeks of its existence in 2022, until it was abandoned under a barrage of criticism for its Orwellian name and unclear mission.

The American Sunlight Project’s goal, as the New York Times reported, is to fight back against “what she and others have described as a coordinated campaign by conservatives and others to undermine researchers, like her, who study the sources of disinformation.” In other words, the newest addition to this expanding bureaucracy is an *anti-anti-disinformation* unit — built to defend the fact-checking fraternity against attacks. Jankowicz has become a pugnacious presence on social media, seemingly offering herself as a public spokesperson for what increasingly looks like a political project.

Determining the extent of the damage done to media properties in recent years by self-appointed disinformation monitors is difficult because their influence on the complex machinery that serves online advertising is hard to measure. It is even unclear which groups’ evaluations are heeded in this murky system.

But I can attest that UnHerd has been substantially affected: Though NewsGuard, another disinformation ratings organization, gives us a trust score of 92.5 percent (five points ahead of the New York Times), the GDI at some point last year mysteriously placed us on their “dynamic exclusion list” of publications that supposedly promote disinformation and should be boycotted by advertisers. As a result, tech giant Oracle, which has a relationship with the GDI, provided a poor “brand safety” rating to our ad agency, and we received only a tiny fraction of the ad revenue the agency had predicted for our audience. Thankfully, we are primarily subscriber-funded, but for smaller publications more reliant on ad revenue, this would be a death knell to their business.

What did UnHerd do to provoke the GDI's disapproval? After repeatedly asking the organization for an explanation, we eventually got an answer: "Our team re-reviewed the domain, the rating will not change as it continues to have anti-LGBTQI+ narratives. … The site authors have been called out for being anti-trans. Kathleen Stock is acknowledged as a 'prominent gender-critical' feminist."

They did not point to any factual errors — their complaint was with the viewpoints of some of our contributors. In addition to decrying Stock, a prominent British philosopher and co-director of the Lesbian Project, the GDI email pointed to Julie Bindel, a lifelong campaigner to stop violence against women, and Debbie Hayton, who is transgender. Apparently the GDI equates "gender-critical" beliefs, or maintaining that biological sex differences exist, with "disinformation" — despite the fact that those beliefs are specifically protected in British law.

The GDI similarly targets other issues — such as climate change and the origins of the coronavirus — that are more properly the subject of robust debate, not matters of "disinformation" if a writer simply has a viewpoint that GDI disapproves.

When the index was originally set up, in 2018, it defined disinformation as "deliberately false content, designed to deceive." On this basis, you could see the argument for having fact-checkers to identify the most egregious offenders and call them out. But mission creep has set in at the GDI. It has since come up with a definition of disinformation that encompasses anything that deploys an "adversarial narrative" — stories that might be factually true but pit people against one another by creating "a risk of harm to at-risk individuals, groups or institutions" — with institutions defined as including "the current scientific or medical consensus."

GDI co-founder Clare Melford explained in a 2021 interview at the London School of Economics how this expanded definition was more "useful," as it allowed the GDI to go beyond fact-checking to flagging any online material the organization deemed "harmful" or "divisive."

In December 2022, the GDI issued a report listing the 10 U.S. publications that posed the most "risk" of promoting "disinformation." It looked distinctly like a list of the country's most-read conservative websites, including the New York Post and RealClearPolitics.

In December last year, two publications on the GDI list, the Daily Wire and the Federalist, teamed up with the attorney general of Texas to sue the State Department for helping fund GDI and NewsGuard. In recent years, GDI has received hundreds of thousands of dollars in backing from the State Department and other government-related entities. The British government is an even heavier backer: From 2019 to 2023, the Conservative government — perhaps to the astonishment of Tory voters, if they had been aware — directed about $3.2 million to the GDI, which also is backed by George Soros's Open Society Foundations and other liberal organizations.

The de facto alliance between government and groups working to defund disfavored publications — a sort of state censorship laundering arrangement — is particularly alarming. Congress is awakening to the problem: It sent a message on this front with the 2024 National Defense Authorization Act, barring the Defense Department from placing military-recruitment advertising in publications utilizing GDI, NewsGuard or "any similar entity."

But as we have seen at UnHerd, the unaddressed problem with these disinformation referees is how their rulings affect online ad services themselves, not just advertisers, with the power to throttle revenue to publications simply for ideological reasons.

Share this articleNo subscription required to readShare

It isn't even clear that the bosses of big tech companies understand the extent that their own organizations have become entangled in this movement. A spokesman for Oracle last year announced that they would be ending their relationship with the GDI on free speech grounds, but our research shows that Oracle is still collaborating via their ad tech platform, Grapeshot. One wonders: Is Oracle's founder and chairman, Larry Ellison, a Republican donor, aware of this? Meanwhile, Elon Musk responded to our investigation by saying on X that the GDI should be "shut down, with recriminations for the miscreants," apparently unaware that his own company, X, is collaborating with the GDI via X's partnership with Integral Ad Science for brand safety information.

There is no doubt that an open, free internet means bad information can travel like never before. But attempts to impose censorship of political speech under the apparently innocuous banner of combating "disinformation" — whether the projects are highly publicized, like the American Sunlight Project, or secretive and pseudo-technical, like the Global Disinformation Index — amount to a much greater risk to a functioning democracy. Not only does censorship not work, but it also adds fuels to the flames of division and paranoia. Next time you hear someone casually use the word "disinformation," be skeptical: They might well be making the problem worse.

# Document 7 - **Is artificial intelligence a threat to journalism or will the technology destroy itself?**

Hitching a struggling media industry to the wagon of AI won't serve our interests in the long run

Samantha Floreani, ***The Observer,*** Sat 5 Aug 2023

Before we start, I want to let you know that a human wrote this article. The same can't be said for many articles from News Corp, which is reportedly using generative AI to produce 3,000 Australian news stories per week. It isn't alone. Media corporations around the world are increasingly using AI to generate content.

By now, I hope it's common knowledge that large language models such as GPT-4 do not produce facts; rather, they predict language. We can think of ChatGPT as an "automated mansplaining machine" – often wrong, but always confident. Even with assurances of human oversight, we should be concerned when material generated this way is repackaged as journalism. Aside from the issues of inaccuracy and misinformation, it also makes for truly awful reading.

Content farms are nothing new; media outlets were publishing trash long before the arrival of ChatGPT. What has changed is the speed, scale and spread of this chaff. For better or worse, News Corp has huge reach across Australia so its use of AI warrants attention. The generation of this material appears to be limited to local "service information" churned out en masse, such as stories about where to find the cheapest fuel or traffic updates. Yet we shouldn't be too reassured because it does signal where things might be headed.

In January, tech news outlet CNET was caught publishing articles generated by AI that were riddled with errors. Since then, many readers have been bracing themselves for an onslaught of AI generated reporting. Meanwhile, CNET workers and Hollywood writers alike are unionising and striking in protest of (among other things) AI-generated writing, and they are calling for better protections and accountability regarding the use of AI. So, is it time for Australian journalists to join the call for AI regulation?

The use of generative AI is part of a broader shift of mainstream media organisations towards acting like digital platforms that are data-hungry, algorithmically optimised, and desperate to monetise our attention. Media corporations' opposition to crucial reforms to the Privacy Act, which would help impede this behaviour and better protect us online, makes this strategy abundantly clear. The longstanding problem of dwindling profits in traditional media in the digital economy has led some outlets to adopt digital platforms' surveillance capitalism business model. After all, if you can't beat 'em, join 'em. Adding AI generated content into the mix will make things worse, not better.

What happens when the web becomes dominated by so much AI generated content that new models are trained not on human-made material, but on AI outputs? Will we be left with some kind of cursed digital ouroboros eating its own tail?

It's what Jathan Sadowski has dubbed Habsburg AI, referring to an infamously inbred European royal dynasty. Habsburg AI is a system that is so heavily trained on the outputs of other generative AIs that it becomes an inbred mutant, replete with exaggerated, grotesque features.

As it turns out, research suggests that large language models, like the one that powers ChatGPT, quickly collapse when the data they are trained on is created by other AIs instead of original material from humans. Other research found that without fresh data, an autophagous loop is created, doomed to a progressive decline in the quality of content. One researcher said "we're about to fill the internet with blah". Media organisations using AI to generate a huge amount of content are accelerating the problem. But maybe this is cause for a dark optimism; rampant AI generated content could seed its own destruction.

AI in the media doesn't have to be bad news. There are other AI applications that could benefit the public. For example, it can improve accessibility by helping with tasks such as transcribing audio content, generating image descriptions, or facilitating text-to-speech delivery. These are genuinely exciting applications.

Hitching a struggling media industry to the wagon of generative AI and surveillance capitalism won't serve Australia's interests in the long run. People in regional areas deserve better, genuine, local reporting, and Australian journalists deserve protection from the encroachment of AI on their jobs. Australia needs a strong, sustainable and diverse media to hold those in power to account and keep people informed – rather than a system that replicates the woes exported from Silicon Valley.

*Samantha Floreani is a digital rights activist and writer based in Naarm*

https://www.washingtonpost.com/opinions/2024/02/06/ai-news-business-links-google-chatgpt/

# Document 8 - **The real wolf menacing the news business? AI.**

By Jim Albrecht, ***The Washington Post***, February 6, 2024

*Jim Albrecht was senior director of news ecosystem products at Google from 2017 to 2023.*

The news publishing industry has always reviled new technology, whether it was radio or television, the internet or, now, generative artificial intelligence. After all, newspapers long had a monopoly on the distribution of information, and each innovation pared back the exclusiveness of that franchise.

The news industry's problem has also been my problem. For the past seven years, I ran a team at Google focused on making the web ecosystem more hospitable to news publishers. We built products to make the production of expensive journalism cheaper (giving them cutting-edge AI document analysis and transcription tools), to make it easier for people to buy subscriptions, and to let publishers showcase their editorial viewpoints and thus find their audiences more effectively. In aggregate, these things delivered billions of dollars of value to publishers around the world.

But they did not fundamentally alter the fact that the internet had hollowed out the value of the daily newspaper. Back in the day, if you wanted to know a sports score, a stock quote, a movie showtime, where the garage sales were or what concerts were coming up, you looked in the newspaper. Now, the web allows you to find this information more quickly elsewhere. So, if consumers once had 20 reasons to buy a newspaper, now they had only one: news — the labor-intensive, expensive work of reporting and writing the news — which isn't a thing advertisers are especially excited to be associated with.

To combat this turn of affairs, news publishers, first in Europe but increasingly around the world, began turning to regulators and legislators to restore their past dominance — or at least their profitability. And I had to figure out how Google would respond to these demands.

The publishers' complaints were premised on the idea that web platforms such as Google and Facebook were stealing from them by posting — or even allowing publishers to post — headlines and blurbs linking to their stories. This was always a silly complaint because of a universal truism of the internet: *Everybody wants traffic!* Just look at the time and money publishers spend putting their links and content on those platforms — paying search-engine optimization companies and social media managers to get more links higher on the page. We found ourselves in the disorienting situation of having one team from a publisher charge, "You are stealing from us by placing our results on your site," while another team complained, "It's critically important to us that you place our results on your site more often and at higher levels of prominence!"

This is not to say that news publishers had no legitimate complaints: Until 2017, Google would rarely link to stories behind a paywall, which was crippling to the subscription model that web publishers were coming to rely on. The selection of news results was imperfect, sometimes placing a site that had done painstaking original reporting below a less authoritative site that had done a quick rewrite of that scoop; and many readers were only interested in scanning the headlines and didn't click to read the actual story. Google fixed the first of these, made steady progress against the second and is powerless to solve the third — a battle that cover designers and front-page editors had been fighting for decades before the web.

In any event, regulators pursued the *illegitimate* complaint: the idea that platforms should pay publishers every time they display a headline/blurb or sometimes even for the act of linking itself. As these regulations or threats of regulation spread around the world — Europe, Australia, Indonesia, Brazil, Canada — I spent more and more time preparing to disable news products, or disabling search, or building accounting systems to count "snippets" and calculate payments. That meant I spent less time giving journalists research and transcription tools, or building mechanisms to help retain subscribers. As for Facebook, each year, its traffic to news publishers plummeted. It is a well-known economic fact that when you take a thing with an established market price and impose a fixed price level above that, demand goes down. Prior to these laws, no one ever asked permission to link to a website or paid to do so. Quite the contrary, if anyone got paid, it was the party doing the linking. Why? Because everybody wants traffic! After all, this is why advertising businesses — publishers and platforms alike — can exist in the first place. They offer distribution to advertisers, and the advertisers pay them because distribution is valuable and seldom free.

While this sideshow was going on, we would hear how much closer large language models (LLMs) had gotten to reproducing human-level composition. Then LLM-based features began to show up in multiple products — grammar checking, autocomplete, etc. — and actually worked. To me, watching publishers bicker about payment for search results while LLMs advanced at a silent, frenetic pace was like watching people squabble about the floral arrangements at an outdoor wedding while the largest storm cloud you can imagine moves silently closer.

And then, like a thunderclap, ChatGPT launched and put everything in stark relief. The problem has never been that platforms post links to news articles — that's what they *should* do. The problem is that new technology has created a landscape where they might not need to link to news sites at all — they can just take the news, have a robot rewrite it and publish it in their own products.

And, for me, the world turned suddenly upside down. The absurd demand of news publishers — “send me traffic and then pay me for having done so!” — would soon be eclipsed by an equally absurd proposition from the tech industry: “How about we build a product on your content and send you little or no traffic in return?” In the long run, neither of these irrationalities can stand. They'll either wither away because of their own economic absurdity or end up in the crosshairs of courts, legislators or regulators.

But having seen firsthand the feckless way in which regulators lined up behind the first of those propositions, I'm bracing myself for how they'll handle the second. The stakes couldn't be higher. On one side of the conflict sits existential risk for the publishing industry; on the other, existential risk for technological innovation.

Share this articleNo subscription required to readShare

First come the courts. The New York Times fired the opening salvo in December in a suit charging OpenAI and Microsoft with violation of its copyright, starting with the use of its documents in training OpenAI's LLMs.

It seems quite plausible that the tech companies will win this first round. AI products transform text into geometric relationships that are fundamentally different from the news stories they came from, and these mathematical “vectors” cannot be substituted for those original stories. In other words, LLMs seem to pass the tests for fair use.

Only when you put an LLM into a consumer product such as a chatbot or search engine do you see it potentially infringing on copyright. An LLM, after all, can produce variations on *any* text. But even then, while those variations very clearly *can* substitute for the originals on which the model was trained, they are indeed *variations* — akin to the sort of human rewrites that publishing companies do all the time. (Note that the Times's recent suit presents evidence of ChatGPT reciting paragraphs of text from Times content — clearly a copyright violation — but this can be easily fixed, just as human rewriters can be trained not to repeat text verbatim from other sources.) Moreover, no one can own a copyright to mere facts. And yet, if one cannot, then how can the rights of content producers be protected?

The answer, I think, lies in the fact that LLMs tend to hallucinate — make up things that aren't real — and that they are so expensive to train that the models are updated on the order of months, rather than days or minutes. As the Times points out in its suit, generative AI products tend to rely on a process known as “grounding,” in which the statements made by the AI are checked against relevant source documents to ensure that the AI is *not* making things up. This process is especially critical if a user is asking about a recent event in which the relevant facts did not exist at the time of the LLM's training. In such cases, the AI can only answer accurately if it retrieves those facts from recent grounding documents. These documents are the essence of the work newspapers do — sourcing and reporting new facts — and the fruits of that labor should reasonably belong to those who perform it.

The courts might or might not find this distinction between training and grounding compelling. If they don't, Congress must step in. By legislating copyright protection for content used by AI for grounding purposes, Congress has an opportunity to create a copyright framework that achieves many competing social goals. It would permit continued innovation in artificial intelligence via the training and testing of LLMs; it would require licensing of content that AI applications use to verify their statements or look up new facts; and those licensing payments would financially sustain and incentivize the news media's most important work — the discovery and verification of new information — rather than forcing the tech industry to make blanket payments for rewrites of what is already long known.

Such legislation would provide publishers new opportunities to generate revenue. If LLM training is indeed held to be a fair use but grounding is not, the publishers' ability to verify the information or infuse it with up-to-date facts becomes not merely valuable but potentially differentiating for their own products. A small, local media company would be able to license its local articles and factual information to generative AI services, but a large media company might choose not to. It might rather offer its subscribers a differentiated AI service of its own, perhaps based on OpenAI or Google APIs, but enriched with proprietary information not available to other providers. Such a service might be more timely, comprehensive and relevant to its subscribers than the tech vendors' own products, and would enable publishers to extend their services back into categories of information they haven't effectively competed in since the print era.

If a court decision or congressional legislation were to rewrite the rules as described, what would the new media world look like? First, to take advantage of the new framework, media companies would need to understand that consumer expectations are about to change dramatically.

In the print era, publishers created "articles," printed them on paper and distributed that paper to their readers. The web changed everything about the distribution and the literal paper, while the articles remained mostly untouched. But in the future, publishers will have to think less about those articles and more about conversations with users. The users will interact less and less with the actual articles and instead talk about the articles with what the tech industry used to call "intelligent agents."

Back in the 1990s, Microsoft introduced Clippy — a simpering, eye-batting paper clip who interrupted you at inopportune moments to ask you whether you needed help. Microsoft put Clippy out of his misery long ago, but as is so often the case, the technology finally caught up to the idea.

The new breed of LLM-powered Clippy is going to do all the things Microsoft hoped it would in 1996: brief you on the news, your day, your emails; respond for you; answer your questions; help with your work. One morning, it might let you know that "The Washington Post announced it has launched a new AI assistant, called Marty." As you ask for more info, it says, "Why don't I just ask him to join us right now since you're a subscriber." Marty joins the conversation and gives you a roundup of The Post's latest coverage, responds to a question you have with a relevant info graphic, updates you on some political gossip and recommends a newly reviewed TV series based on your interests. (Because you're a subscriber, he knows what you like.) "Can you find me a restaurant for Thursday night?" you ask, and Marty gives you some of the best local options and what they're known for, and he notes that he can offer you a discount at one of them. Maybe you decide to make Marty a part of your daily briefing or, on the other hand, maybe you turn to your ChatGPT agent and ask, "So what do I need you for?" She might say, "I can do things like make travel arrangements," to which Marty responds, "We have a travel agent we work with, as well. Shall I ask ExpediaBot to join?" Welcome to your new daily newspaper.

The details could turn out very differently, of course. It depends on the outcome of these current copyright disputes and on the ability of publishers to envision a future that looks very different from their past. But one thing is certain: As with the web 30 years ago, those details will determine whether the news business reclaims its status as the premier vendor of reliable information or falls into a final, unrecoverable decline.

## Disinformation is on the rise. How does it work? – The Article

Understanding it will lead to better ways to fight it

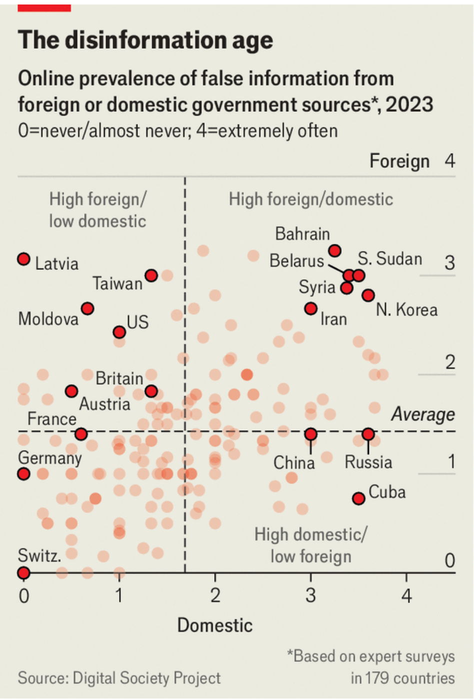
***The Economist***, May 1st 2024

Listen to this story.  Audio version to be found on Cahier de Prépa

In January 2024, in the run-up to elections in Taiwan, hundreds of video posts appeared on YouTube, Instagram, X and other social platforms entitled “The Secret History of Tsai Ing-wen”. News anchors, speaking English and Chinese, made a series of false claims about Ms Tsai, the outgoing president, and her ruling party. On election day itself, January 13th, an audio clip began to circulate in which Terry Gou, a candidate who had dropped out of the race in November, seemed to endorse the candidate of the China-friendly KMT party (in fact, Mr Gou made no endorsement).

Both the video clips and audio were probably created using artificial intelligence (AI) and posted by a Chinese state-backed propaganda group known variously as Spamouflage, Dragonbridge and Storm-1376. In a report released on April 5th, the Threat Intelligence team at Microsoft, a tech firm, said this was the first time it had seen a nation-state use ai-generated material to sway a foreign election.

The news anchors in the videos were made using CapCut, an app made by ByteDance, the Chinese parent company of TikTok. At their peak, the videos were being shared 100 times a minute, but were swiftly identified and taken down. Overall, few people probably saw them. But China is likely to be using Taiwan as a testbed for ideas it plans to deploy elsewhere, a Taiwanese official told the *Taipei Times*. Taiwan’s election is a sign of things to come, as ai supercharges the production of disinformation (that is, information that is intended to deceive). The country’s social media are flooded with one of the world’s highest levels of disinformation coming from foreign governments (see chart). American social media are not far behind on the same measure.



The disinformation age

Online prevalence of false information from foreign or domestic government sources*, 2023

0=never/almost never; 4=extremely often

Source: Digital Society Project

*Based on expert surveys in 179 countries

chart: the economist

In a year when half the world is holding elections and new technology is making it easier than ever to make and spread disinformation, the need for governments, companies and individuals around the world to grapple with the problem has never been more urgent. Security experts are raising the alarm too—more than 1,400 of them recently told the World Economic Forum that disinformation and misinformation (incorrect information that is shared unwittingly) were the biggest global risks in the next two years, even more dangerous than war, extreme weather or inflation.

The fog of information war

Much is still murky, including how much disinformation there is and exactly how (and how much) it shapes opinions and behaviour. Nevertheless, researchers are beginning to understand how disinformation networks operate and are developing ways to identify and monitor them in the wild. Some countries—from Taiwan and Sweden to Brazil—have implemented policies to fight the problem, which could provide useful lessons for others.

This special section will first explain how disinformation campaigns work. It will consider the role of ai—both negative (creating disinformation) and positive (detecting and mitigating it). And it will assess the emerging tools and policies aiming to fight back against the problem.

Bad information can take many forms and affect many fields. For many years the standard way of doing things, says Amil Khan, a former journalist who now studies disinformation, was to use hundreds or thousands of social-media accounts, controlled by a single piece of software, to pump out the same message or link, or to "like" or reshare particular posts. On a large scale, this "co-ordinated inauthentic behaviour" (CIB) can fool the curation algorithms on a social network such as Facebook or X into thinking there is a groundswell of interest in, or support for, a particular viewpoint. The algorithm then promotes those posts to real users, who may then share them with their own followers.

One example of CIB analysed by Mr Khan's firm, Valent Projects, targeted Daewoo, a South Korean company. After Daewoo won a construction contract worth $2.6bn from the Iraqi government, it was attacked by Chinese disinformation networks, which spread false stories about the company in an effort to get its contract cancelled and awarded to a Chinese company instead. Daewoo was said to be a front for a Western plan to exploit Iraq's resources; made-up comments attributed to American officials were cited as evidence that America was trying to undermine Chinese-Iraqi relations. Such claims were debunked by fact-checking outfits based in Iraq and Qatar, but that did little to hamper their spread.

CIB operations are fairly easy to spot. Meta, the tech firm behind Facebook, now finds and shuts down such networks quite quickly, says Mr Khan (though X is slower off the mark, he adds). In its most recent report on CIB shutdowns, released in February, Meta describes three such operations, in China, Myanmar and Ukraine. The report notes that early CIB networks targeting Ukraine were simple and could be traced to Russian intelligence services.

Since 2022, however, disinformation campaigns have taken on a new form. Run by "deniable entities" such as marketing companies or troll farms without direct state links, they post on a range of social networks and blogging platforms and create entire fake websites. Since May 2023 the number of ai-generated news outlets peddling misleading information has risen from 49 to 802, according to NewsGuard, an American organisation that monitors disinformation. These sites mostly feature innocuous articles, generated by ai, but with disinformation mixed in.

An example is *dc Weekly*, an apparently American website that was central to furthering the Russian-led disinformation campaign alleging that Olena Zelenska, Ukraine's first lady, had spent $1.1m on a shopping spree on New York's swanky Fifth Avenue (see interactive). That story, which was tracked by researchers at Clemson University, began with a video on YouTube, passed through several African news websites and an ai-generated site before being planted on social media and boosted by Russian propaganda outfits. It was shared 20,000 times on X.

Mr Khan calls the accounts and sites that plant the story "seeders". Rather than using hundreds of fake accounts to promote these sites' material, distribution instead relies on a few so-called "spreaders"—social-media accounts with large numbers of followers. Spreader accounts typically build a following by posting about football, or featuring scantily clad women. "And then they'll flip," says Mr Khan: they start mixing in disinformation from seeders, by linking to or reposting their content. Meta's threat report from November 2023 noted that it had seen the Chinese outfit Spamouflage/Storm-1376 operating on more than 50 platforms, "and it primarily seeded content on blogging platforms and forums like Medium, Reddit and Quora before sharing links to that content on [our platforms]".

In poor countries with few opportunities for young, tech-savvy men, there is a cottage industry of building up spreader accounts and then selling them to malicious actors once they reach 100,000 followers, says Mr Khan. The challenge with identifying spreaders is that their behaviour is genuine, at least to begin with, and they are not the originators of disinformation, but merely distributors of it. Spreader accounts may continue to post about other things, with disinformation mixed in every so often, to avoid detection.

Valent has seen this more sophisticated approach being used to spread disinformation from Russia in European countries, and to promote hard-right material in Britain. In the latter case, the spreader accounts used gossipy posts about the British royal family to attract followers, before flipping to political propaganda about how "low-traffic neighbourhoods" (areas where through-traffic is discouraged) are a globalist plot. Similarly, Microsoft has detailed Storm-1376's use of this newer distribution model to spread disinformation about wildfires in Hawaii (supposedly started by an American "weather weapon"), to amplify criticism of the Japanese government in South Korea, and to stoke conspiracy theories about a train derailment in Kentucky in November 2023.

When many accounts are using exactly the same wording, spotting CIB is relatively simple. It is not unusual for narratives to suddenly “trend” on social media, but a spike in mentions of a particular topic in an array of different languages, or posted by accounts seemingly scattered around the world, might be a hint that foul play is involved. Similarly, disinformation hunters can examine clusters of accounts that push a similar message. Dodgy accounts may all have the same date of creation, the same number of followers or the same ratio of followers to following (because they have bought fake followers in bulk in a bid to look authentic).

But spotting seeders and spreaders under the newer distribution model is more difficult. They are propagating a particular narrative, but the seeder articles and posts, and the spreader posts that promote them, may all use different wording. Valent is trying to solve this problem using ai: its system, called Ariadne, takes in feeds from social platforms and looks for common underlying narratives and sentiments to spot unusual, co-ordinated action. Unlike previous approaches based on keywords, “the latest models let us do work on sentiment that we couldn’t do before”, says Mr Khan.

Another way to identify spreader accounts is described in a recent working paper from Brookings, a think-tank based in Washington, dc. Maryam Saeedi, an economist at Carnegie Mellon University, and her colleagues analysed 62m posts from X written in Farsi, relating to the wave of anti-government protests in Iran that began in September 2022. The analysis focused on spreader accounts (the researchers call them “imposters”) that started off by pretending to be on the side of the protesters, but then flipped to posting disinformation discrediting the protests.

The researchers began by identifying several hundred imposter accounts by hand. They then trained a classifier algorithm to identify more imposters with similar characteristics, including their posting activity, the pattern of their followers, their use of particular hashtags, how recently the account was created, and so on. The researchers were then able to replicate the identification of these imposter accounts, with 94% accuracy, through network-analysis alone—ie, by scrutinising only their relationship to other accounts, rather than the content of their posts.

This suggests, the researchers say, that it is possible to identify “accounts with a high propensity to engage in disinformation campaigns, even before they do so”. They suggest that social-media platforms could use these kinds of network-analysis methods to calculate a “propaganda score” for each account, made visible to users, to indicate whether it is likely to be a source of disinformation. The imposter-detection algorithm could be further improved, the researchers suggest, using more advanced forms of ai such as natural-language processing and deep learning.

Technology firms and intelligence agencies are no doubt already doing this kind of analysis, though they are understandably reluctant to share details about their methods. Meta says only that it has used ai in its “integrity systems” for many years to protect users and enforce its rules. But both academics and civil-society groups say that no single method can be used to automatically detect all disinformation—the tactics used are often bespoke to specific campaigns and rely on human analysts to check the results and provide interpretation and nuance.

ai can, however, help in a different way: by spotting deceptive content directly, through analysis of individual posts, articles, sound clips or videos. DARPA, the special-projects research arm of America’s Department of Defence, has been funding research into “detecting, attributing and characterising manipulated and synthesised media” as part of its “semantic forensics” programme, to create a toolbox of defences. In March it published an open-source repository of several of the projects it has funded, with links to downloadable source code, and announced a series of “spot the deepfake” challenges. Its aim is to encourage academic and commercial users to combine, improve and ultimately deploy these tools, all of which rely on ai in some form, says Wil Corvey of DARPA, who manages the programme.

Although a single analytic tool may not always be reliable, he says, combining several of them can greatly improve accuracy. Consider, for example, the problem of working out whether a video of a politician is genuine or not. Using authentic data of the person in question, it is possible to train an ai model that learns their characteristics, such as patterns of head tilt or facial movements while speaking. This can then be used to analyse a suspected deepfake for authenticity. Better still, explains Dr Corvey, it can be combined with other techniques, such as heartbeat detection from video, which is difficult to fake. (Heartbeats can be spotted by looking for tiny variations in skin colour, particularly on the forehead.) Other tools in DARPA’s catalogue are capable of spotting ai-generated text, synthesised or edited images, and deepfake audio and video. The results for fake-audio detection are “particularly robust”, Dr Corvey says. And a fake-audio track can, of course, indicate that the accompanying video is also fake.

DARPA’s is not the only effort of this kind. Oren Etzioni, a computer scientist and former head of the Allen Institute for Artificial Intelligence, a research outfit, founded TrueMedia.org, a non-profit group, in

January to expand access to detection tools. On April 2nd the group unveiled a website which brings together open-source and commercially available tools, through a free web interface, to provide a one-stop-shop for detection of synthetic or manipulated images, video and audio using multiple tools simultaneously.

When it comes to bad content, ai is both a sword and shield, notes Nick Clegg, head of global affairs at Meta. For his part, Dr Corvey says he is optimistic that defensive ai-powered detection tools can stay ahead of offensive generation tools. Dr Howard, of Oxford University, agrees—at least for now. This is a “lucky moment” in which technology firms can spot fake videos pretty reliably, he says, “though I don’t know that this is going to last for ever.”

Renée DiResta, who studies information flows at the Stanford Internet Observatory, is less convinced. Today’s detection tools may work well in a controlled environment when there is plenty of time to make an assessment, she says. But when it comes to making snap judgments in the heat of the moment, “I don’t think the defender is necessarily favoured.” Besides, she observes, there is a much deeper problem. Even if deceptive media can be detected with perfect accuracy, not everyone will believe that a fake video is fake. She cites the example of fake audio clips that went viral just before an election in Slovakia in September 2023, in which a politician was apparently heard discussing election-rigging with a journalist. He later lost the election. “People are highly resistant to fact-checks if they don’t like the fact checker,” she says. That means the mitigation of disinformation will require much more than just technology. ■