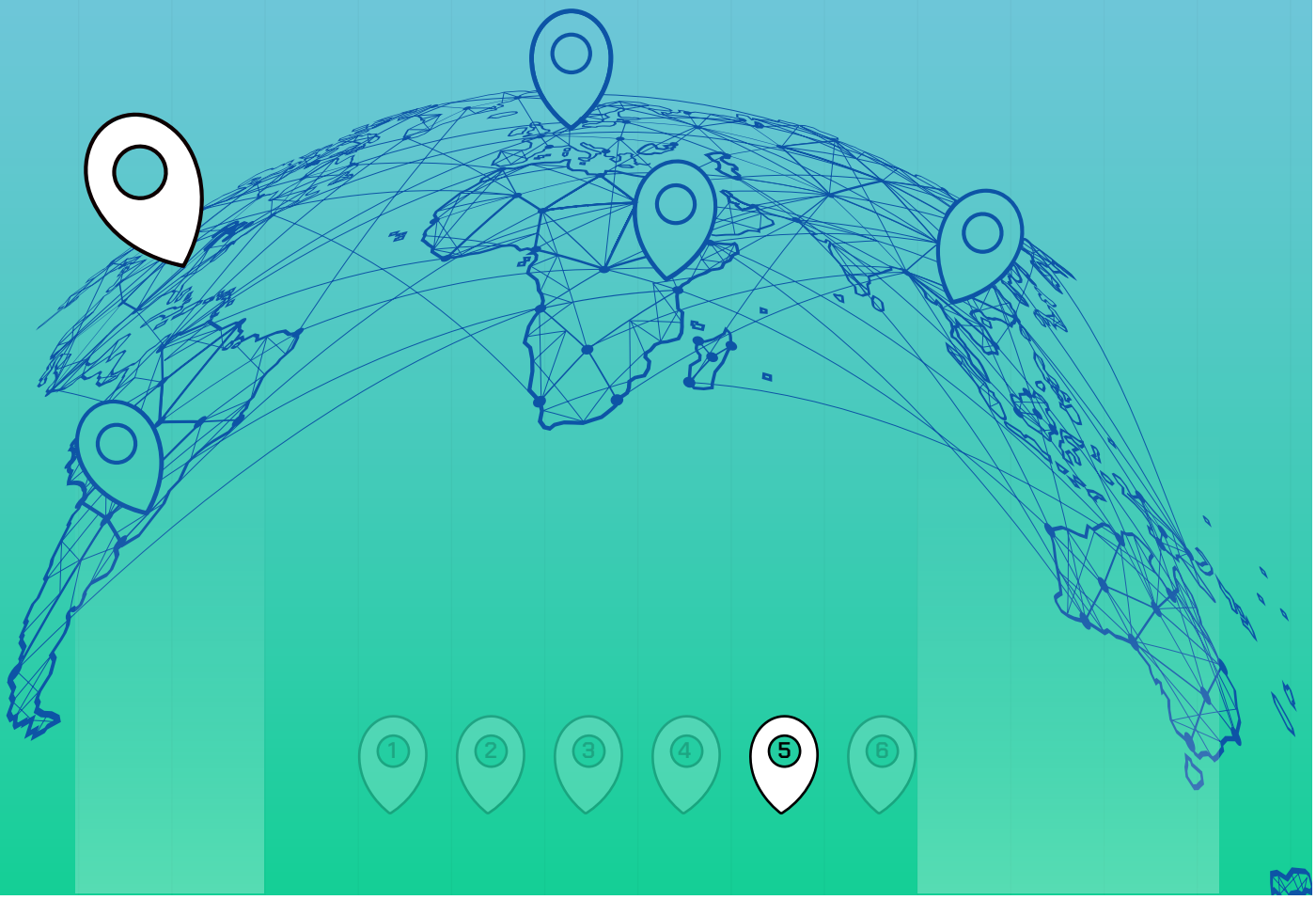
upgrade
democracy

Research Series: Reinhard Mohn Prize

Countering disinformation in the United States

Shwetha Rao



Bertelsmann FOUNDATION

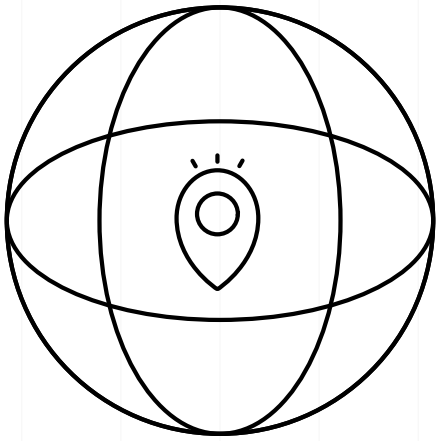
BertelsmannStiftung

upgrade
democracy

Research Series: Reinhard Mohn Prize

Countering disinformation in the United States

Shwetha Rao



This report is part of an international research series on "Strengthening Democracy, Countering Disinformation."

Preface

Dear Reader,

In today's digital world, countering disinformation has emerged as an essential endeavour to uphold democratic values worldwide. While there is a shared understanding that concerted efforts from various stakeholders and at different levels are needed to address this issue, we still lack a comprehensive understanding of the strategies and initiatives in place, let alone their impact and how to accurately measure it.

As part of the **Reinhard Mohn Prize** – 'Strengthening Democracy, Countering Disinformation', we sought to illuminate the way forward by identifying exemplary models and innovative approaches to countering disinformation around the world. Our goal was to gain insight into the where, how, and why of disinformation, and to respond accordingly. Across the globe, there are countless successful and impactful examples of individuals, initiatives, and organisations dedicated to countering disinformation. Our aim was to learn from them and empower us all to learn from each other.

This series of six reports covering Africa, Asia-Pacific, Europe, North and Latin America, and a global overview of government responses to disinformation, consolidates our findings in the hope of providing you not just with key insights, but also with actionable recommendations. These reports couldn't be clearer: We can all learn from each other. From the technology enhanced fact-checking approaches of **Chequeado** (Argentina) or **Aos Fatos** (Brazil), to the community-driven debunking of **JamiiCheck** (Tanzania) or the rapid response mechanism at **Real411** (South Africa), to the thought-provoking media literacy trainings by **Fact Shala** (India) and **Mafindo** (Indonesia) – there is so much knowledge out there that we could write entire books about it.

We invite you to explore, learn, and be inspired. Because there is hope for a healthier information ecosystem thanks to the efforts of everyone we encountered.

Finally, we would like to express our deepest gratitude to the outstanding authors of these reports, as well as to all the experts who participated in our workshops in Nairobi, Bangkok, Buenos Aires, Washington D.C., and Brussels. It is your expertise and your dedication to strengthening democracy – regardless of the challenges faced – that have made this series so insightful and special.

Our warmest,

Cathleen Berger Charlotte Freihse

Contents

Preface 3

Introduction 5

1. The U.S. landscape 7
   - 1.1 The lack of legislation: Section 230 7
   - 1.2 Technology companies: Lacking accountability 8
   - 1.3 Erosion of local media 9
   - 1.4 From disinformation to online radicalisation 11

2. Case studies and solutions 13
   - 2.1 Disinformation solutions: Pre-bunking (Jigsaw) 13
   - 2.2 Disinformation solutions: De-platforming and demonetising 14
   - 2.3 Extremism solutions: Hash-sharing 16
   - 2.4 AI and Natural Language Processing 18
   - 2.5 AI deepfake identification: Adobe and Content Authenticity Initiative 19
   - 2.6 Media literacy: The Trust Project 20

3. Conclusion 22

References 23

Publishing credits & legal notice 29

Introduction

One of the most pernicious threats to American democracy, and indeed to democratic institutions around the world, is the electorate's inability to discern truth from fiction. And while the current dynamic feels particularly acute as the 2024 US presidential election looms on the horizon, dealing with disinformation in the political ecosystem of the United States is nothing new, and its effects often have a contagious effect on elections around the world. In fact, the use of disinformation can be traced back to the Founding Fathers, who used it to rally the colonies against the British Crown. One famous example comes from the less noble exploits of Benjamin Franklin who pushed stories into the public domain implicating King George III's alleged alliance with native Americans, to sway the sentiment of loyalists in the early days of the Republic.

The truth is that the tradition of American journalism since its founding has tended towards sensationalism and bias, rather than the halcyon days of Walter Cronkite, Peter Jennings, and Tom Brokaw. The line between truth and fiction has at times been blurred, but in today's information environment it has all but disappeared. In the vacuum left behind is a toxic stew of disinformation ranging from the politically polarising to the overtly racist and misogynistic. As a result, American democracy is teetering on the brink of destruction, with a population deeply divided over its future and no longer seeing their fellow Americans as political opponents, but as enemies.

It is against this backdrop that we seek to tackle the scourge of disinformation. The stakes are enormous, and the tools available for rapid remedy are frustratingly lacking. With this in mind, the Reinhard Mohn Prize Washington, DC Roundtable gathered on 4 March, 2024 with a sober view of the prospects for surfacing easy solutions, and an understanding that the best contribution the group could make to this ongoing challenge would be to dig deeper into the problem, and highlight some best practices that are moving the needle towards a better-informed electorate. These conversations were part of a research series that informed the direction and content of this report.

To that end, this report begins by exploring the legal framework surrounding disinformation in the social media era, with a particular focus on Section 230 of the Communications Decency Act. It examines the impact of the immunity that big tech companies maintain to protect them from disinformation and other harmful material that users may post on their respective platforms. From there, we move into a related area that looks at the profit motive and incentives involved in spreading disinformation, and whether it is possible to put the proverbial toothpaste back in the tube.

The rise of social media has also indirectly affected an industry that was once a bedrock of information for the average citizen: local news. As advertising revenue shifts to digital media and newsrooms continue to shrink as a result of widespread cost-cutting, local media has been a major casualty of this information revolution. Without a viable business model to sustain local journalism, citizens are left to sift through a mix of trolls, bots, and purveyors of falsehoods.

With the battlefield ceded to those without journalistic integrity or even a basic sense of decency, it is no wonder that the information space has become fertile ground for radicalisation and political extremism. Our section on this topic in this report looks at the methods and effects of algorithms that push users towards their worst instincts.

If you stop reading there, you might be led to believe that all is lost; it is not. The second part of this report highlights a number of solutions that are directly related to the challenges we've dissected. The first of these solutions is the notion of pre-bunking, or inoculating users from exposure to disinformation. The second is an assessment of how to deal with the disinformation economy through tools such as demonetisation and deplatforming. Third, we offer a forensic tool called 'hash sharing' to document and purge digital platforms of the most damaging content. Fourth, we delve deeper into technological solutions that use artificial intelligence and natural language processing (NLP). Essentially, how to fight fire with fire by using machine learning to identify and remove harmful content. Fifth, we look at the next wave of technology that is likely to impact the upcoming elections in the form of deep fake technology. Adobe's Content Authenticity Initiative offers a potential way forward to address the problem of what happens when you can't trust your own eyes and ears. And finally, the Trust Project brings us back to where many of the conversations about disinformation begin: media literacy. How can we better equip users to separate fact from fiction in an environment that makes that assessment increasingly difficult? This project's six factors can move us closer to achieving that goal.

In the following pages, you may find yourself at times despondent and at times optimistic. Unfortunately, this is more a sign of the times than of the author's outlook. There are reasons for hope and reasons for despair, sometimes in equal measure. Ultimately, we hope is that you will finish reading this report more informed than when you started. From there, it will be up to you to decide what to do with this new-found information.

1 The U.S. landscape

1.1 The lack of legislation: Section 230

In the debate over online disinformation, the First Amendment plays an integral role in the existing landscape and how the future of social media content moderation may evolve. The constitutional protections enshrined in the First Amendment inspired Section 230 of the Communications Decency Act (O'Hara and Campbell 2023). Enacted in 1996, this section has played a significant role in shaping the online legal framework. Under this law, social media platforms are generally immune from liability if their users post illegal content, and are not required to remove such content. Platforms are free to engage in content moderation as they see fit, provided they act in good faith.

The debate over Section 230 has intensified in recent years as the spread of disinformation has led to real, dangerous consequences such as potentially inspiring citizens to inject disinfectants like Clorox and Lysol to treat COVID-19 (Rivera et al. 2020). Pro-moderation advocates believe that Section 230 needs to be repealed, or at least amended to include liability for social media if their platforms allow disinformation or hate speech to flourish. Their argument is based on the belief that false content and hate speech could pose a *"clear and present danger"* to society, and is therefore not protected by the First Amendment (White). Proponents point to high-profile examples of harm resulting from a lack of proper content moderation, such as the Buffalo shooting in 2022, in which a gunman murdered 10 African-Americans in a New York supermarket and live-streamed the attack. The shooter had frequented obscure online chat rooms, some of which had even lower standards of content moderation than mainstream social media platforms. One such platform, the Gab online network, was reported to have no content moderation at all (Stocking et al. 2022). These smaller social platforms often become isolated breeding grounds for bigoted speech and political conspiracy theories. The Buffalo shooter had maintained an active online presence, frequenting the 4chan message board 'Politically Incorrect' (Buffalo Shooting Online Platform Investigative Report). The shooter promoted white supremacist beliefs and conspiracy theories – particularly the 'Great Replacement' theory. He once wrote: *"Every time I think maybe I shouldn't commit to an attack I spend 5 min of [sic] /pol/, then my motivation returns."* Clearly, his online activities guided his actions in the real world. In the wake of such violent repercussions, many government officials have advocated eliminating or weakening the immunities afforded by Section 230 protections. In a recent Senate hearing on child safety online (Rosenblatt et al. 2024), Senator Amy Klobuchar (D-MN) noted, *"It's been 28 years since the internet. We haven't passed any of these [content moderation requirements] bills ... The reason they haven't passed is because of the power of your companies, so let's be really, really clear about that. What you say matters. Your words matter."*

As the debate over online content evolves, some politicians are increasingly defending the open dialogue of the internet as a cornerstone of democratic ideals. Senator Ron Wyden (D – OR) stated: *"Keeping the internet free from government censorship and oppressive government surveillance is critical to free expression and political freedom around the world."* (Blackburn, Menendez Lead Effort to Protect Free Speech Online, 2020) Much of the criticism of taking

action against disinformation stems from the perceived targeting of conservative ideals. For example, Missouri Attorney General Andrew Bailey filed a lawsuit in 2022 alleging that the federal government and technology companies were colluding to unfairly remove conservative viewpoints from social media (Missouri Attorney General Andrew Bailey Obtains Court Order Blocking the Biden Administration from Violating First Amendment, 2022). The case reached the Supreme Court in early 2024. The lawsuit focused on the Biden administration's supposed coercion of social media platforms to remove alleged disinformation surrounding the COVID-19 pandemic or the 2020 presidential election. Bailey called the censorship *"the biggest violation of the First Amendment in our nation's history."* The perception that content moderation is simply a method of suppressing political beliefs is leading to legislative gridlock. The battle for control of the information space is being hampered both by conservative politicians afraid of censorship, and by technology companies that refuse to be accountable.

1.2 Lack of accountability: Technology companies

Technology companies and social media platforms often shirk their responsibility to block or even address false content circulating online. Most social media companies devote some content moderation policies to addressing disinformation, but most such policies are riddled with vague definitions of false content and unclear consequences for users who violate their terms. Facebook's content moderation policy for false information is to *"reduce its distribution by showing it lower in the News Feed,"* but does not itself mandate the removal of the content (False News, 2024). In fact, Facebook is one of the largest platforms for sharing false content. Researchers at New York University found that in 2020, the first year of the pandemic, received more attention than real news and up to six times more engagement (Place 2021).

In fact, user engagement with false or offensive content seems to be a key motivation for social media platforms to continue dragging their feet on content moderation. First and foremost, higher engagement leads to higher user numbers, which means more revenue for the company. Search engine algorithms then amplify posts and accounts that generate engagement, regardless of the veracity of the content. In Instagram's algorithm report, Instagram head Adam Mosseri explains the method behind the platform's Explore feature, which shows images from accounts that certain users may be interested in (Mosseri 2023). The Explore feature's algorithm prioritises popular posts based on their number of likes, comments, and shares. In parallel, sensationalist content is associated with higher levels of engagement, as tested by researcher K. Ali and others in *"Viruses Going Viral: Impact of Fear-Arousing Sensationalist Social Media Messages on User Engagement"* (Ali et al. 2019). The researchers tested 800 Facebook posts about the Zika virus in 2019, measuring user engagement. They found that user engagement *"increased significantly as the level of fear-arousing sensationalism increased from low levels."* As a result, when users engage significantly with inflammatory and false content, it can be amplified by the social media platform.

One example of platforms taking advantage of sensational content is Facebook's controversial 'XCheck' programme. In 2021, the Wall Street Journal published an expose about Facebook's programme, which allows millions of powerful individuals, including politicians and celebrities, to be exempted from some or all of its content moderation guidelines (Horwitz 2021). The Journal found that in 2020, 5.8 million people were on the XCheck programme, which allowed

them to post content that was explicit, inaccurate, or even dangerous. For example, Brazilian footballer Neymar Junior was able to post explicit photos of a woman who had accused him of sexual assault. Facebook had responded to the Journal's investigation by saying that the purpose of the XCheck programme was to *"create an additional step so we can accurately enforce policies on content that could require more understanding,"* rather than to amplify content to increase user engagement.

After the expose was published, Facebook asked its Oversight Board to provide an advisory opinion on its content moderation policies and XCheck, which took over a year to produce (Newton and Barclay 2022). Board members said that Facebook was deliberately reluctant to report facts and figures on the issue, and even when it did provide statistics, it only gave limited aggregate data. The damage caused by this programme may be even greater than previously thought, as Facebook refuses to disclose all the information, citing *"privacy concerns."* However, the Board was able to establish that Facebook took an average of 12 days to make a decision about XCheck members who posted content that violated its policies. In one case, the team took more than 222 days to review a post, allowing hate speech, explicit content, and misinformation to run rampant in users' feeds. Because the incentives to allow sensational content to outweigh the existing consequences, Section 230 immunises technology companies and leaves social media platforms largely unchecked.

1.3 Erosion of local media

As social media and online technology continue to reshape the landscape of news and journalism, communities across the US are grappling with the loss of local news outlets. Once pillars of information and connection, local media outlets, including newspapers, radio stations, and television programmes, are dwindling in numbers. The rise of online journalism, combined with media conglomerates such as Fox News or CNN, has led to local media outlets being bought out or losing airtime. The Local News Initiative found in 2023 that 'local news deserts' are becoming a serious problem in the US, including 204 counties that currently have no local newspapers, digital sites, or radio newsrooms, and another 228 counties that are at risk of becoming news deserts in the coming years (The State of Local News 2023). At the same time, public trust in the news media has plummeted; a Gallup poll in 2023 found that only 32 percent of Americans had a great deal or fair amount of trust in the media, a record low (Younis and Evans 2023). The loss of local media represents a fundamental shift in the way Americans consume information, as local issues are often sidelined in favour of broader national events or international affairs that may never affect individual citizens. In addition, the media giants that dominate television and radio have become more polarised, crowding out independents, centrists, or even audiences hoping to be educated by authentic, unbiased journalism. The Knight Foundation found in a 2022 study that across the political spectrum (polling Republicans, Independents, and Democrats), all groups trusted local news organisations significantly more than the national media (Fioroni 2022). The decline of local media poses serious challenges to community cohesion and democracy at all levels of government.

First and foremost, local media bring economic benefits to their communities, enabling job creation and advertising revenue that benefits both the media outlet and the surrounding businesses. As a result of consumers migrating to digital media platforms for information,

"digital media accounts for well over half the ad spend allocated by marketers, and growing," with the largest beneficiaries being Google and Facebook (Adgate 2021). As a result, local TV stations are losing an estimated $1.873bn to Google Search and Facebook news feeds. Because these technology platforms have exclusive control over their algorithms and subsequent monetisation, local broadcasting stations are not being adequately compensated for their online content, further depriving local media of revenue and further shrinking their newsrooms.

Recently, the US government has attempted to reduce technology companies' monopolisation of content revenue. One example is the proposed Journalism Competition and Preservation Act, which would allow local media to negotiate for fair compensation from Big Tech that profits from their content (Senate Judiciary Committee Advances Bipartisan Bill to Preserve Strong, Independent Journalism and News Organisations, 2023). Senator Amy Klobuchar (D-MN) introduced the bill, stating: *"To preserve strong, independent journalism, news organisations must be able to negotiate on a level playing field with the online platforms that dominate news distribution and digital advertising."* (Cassidy, Klobuchar, Kennedy Introduce Bill to Save Local Journalism, 2023)

Another serious blow to local communities is the loss of ethnic media publications, such as digital newsletters or radio broadcasting stations that cater to different ethnic groups. These forms of local media are an essential pillar of democracy in the US, providing specialised support and information that mainstream news media cannot deliver to marginalised groups. As the US population landscape becomes more and more diverse, these groups need opportunities to participate in community. In fact, the US Census Bureau reports that the country will become 'minority-white' by 2047 (Vespa and Medina 2018). Therefore, ethnic publications will become increasingly critical. While these forms of media often face similar challenges to traditional local media, they also faced unique challenges during the pandemic-related economic downturn in 2020. The Local News Initiative found that *"more than a quarter of the community-based ethnic news outlets documented in a 2020 survey by the University of North Carolina have closed, leaving only 723 still actively producing news."* (Abernathy 2023)

In response, some organisations have founded initiatives to empower and uplift ethnic publications. For example, the Knight Foundation's BloomLab works with Black/non-white focused media to implement more digital technologies to drive revenue and audience growth. One example of their success is the Sacramento Observer, a weekly newspaper based in Sacramento, California that focuses on amplifying Black and LGBTQ+ voices. The paper's publisher, Larry Lee, grew his team from 2.5 full-time employees to 14 in less than two years after working with BloomLab.

More attention needs to be paid to local media to avoid more communities becoming news deserts. Legislation is needed to empower local media against Big Tech to avoid bankruptcy, and independent organisations need to focus on ethnic media publications to empower minority groups across the country.

1.4 From disinformation to online radicalisation

Disinformation and misinformation create confusion and undermine reliable information; the current digital landscape has in turn become a catalyst for the growth of extremist content. Tackling right-wing extremism is complicated by the ambiguity of the First Amendment's protection of free speech. Freedom of speech is restricted when it could pose a threat to public safety, as in the case of threats or hate speech. However, due to the lack of specific language in the First Amendment, the legality of some forms of speech is largely left to judicial interpretation. As a result, legislation such as House Bill 1333, introduced by Representative Bill Ramos (D-WA), has received much pushback (Washington State Legislature, 2024). The bill, which would establish a Domestic Violent Extremism Commission, has raised many concerns about freedom of speech. Liv Finne of the Washington Policy Center, speaking about the bill, told Fox News: *"Speech is not violence. Violence is violence," said Finne. "To equate the two is dangerous and wrong."* (Lambert 2023) American jurisprudence has generally held that unless individuals cross the threshold of participating in or inciting violent activity, they are generally protected under the First Amendment – especially online.

The online landscape has become an integral part of the radicalisation and mobilisation trajectory of extremists. Before the rise of online communication methods, radicalisation groups tended to organise locally, relying on word of mouth or printed content to inspire others to join their cause. But with the growth of digital technology and social media, the landscape of extremism has transformed into a global network, reinforcing the belief in online echo chambers and filter bubbles, an *"algorithmic bias that skews or limits the information an individual user sees on the internet."* (What is a filter bubble?) Where once a lone radical actor might have been isolated from a larger community, now they can reach thousands of like-minded people online in seconds. The ability to use social media is an invaluable skill for terrorist groups hoping to find new recruits. An Islamic State defector, quoted in a 2018 article by the Center for Strategic and International Studies noted the reach of online forums (Jones 2018): *"The media people are more important than the soldiers… They have the power to encourage those inside to fight, and the power to bring more recruits to the Islamic State."*

In the world of domestic extremism, conflicting perspectives complicate lawmakers' ability to enact change. One of the most common conflicts concerns the significance of right-wing extremism in the United States. Right-wing extremism, including the Buffalo shooting, often centres on racial or ethnic supremacy and resistance against governmental authority. Meanwhile, the far left focuses on anti-capitalist, anti-authoritarian, and environmental ideologies. For years, academics and researchers have noted the growing prevalence and danger of right-wing extremist violence, but the US government has been slow to recognise this threat. 'The Rising Threat of Domestic Terrorism', a 2021 report by the US Senate Committee on Homeland Security & Government Affairs *"acknowledged that white supremacist violence is one of the most potent forces driving domestic terrorism."* (Haugen 2022) The Senate Committee report went on to reveal that the Department of Homeland Security (DHS) had previously reported its concerns about right-wing extremism in a non-public report in 2009, but that the assessment, *"while accurate, was ultimately rescinded by DHS due to political pressure."*

The US government's previous lack of attention to right-wing extremism can be attributed to a longstanding focus on Islamist/jihadist terrorism. Since the 9/11 attacks, during the presi-

dency of George W. Bush, the Bush administration and subsequent presidential administrations have emphasised the Global War on Terror, seeking to prevent terrorist groups in the Middle East from infiltrating or attacking the United States (Global War on Terror). Other forms of terrorism, including right-wing, left-wing and single-issue extremist groups, have been sidelined, while Muslim Americans have complained that they have been unfairly targeted. The FBI reported that hate crimes against Muslims increased seventeenfold in 2002, the year after 9/11 (UNITED STATES, 2023). Following the terrorist attacks against Israel on 7 October 2023, the Council of American-Islamic Relations reported a 180 percent rise in complaints of Islamophobia and discrimination in the United States when compared to 2022 (Singh 2024). At the same time, the Anti-Defamation League reported that white supremacist propaganda efforts were unprecedented, reaching the highest levels ever recorded in 2022 (White Supremacist Propaganda Soars to All-Time High in 2022, 2022). The ADL recorded a 38 percent increase in one year, with antisemitic propaganda more than doubling. While all forms of extremism must be addressed swiftly and appropriately, prioritising one form can allow others to flourish.

2 Case studies and solutions

2.1 Anti-disinformation solutions: Pre-bunking (Jigsaw)

One might assume that the solution to false content is simple – refute fraudulent claims with facts. However, studies have shown that once audiences are exposed to false content, they are much less likely to believe contradictory information (regardless of its truth). In fact, a 2012 study on misinformation published by SageJournals concluded that even after learning that a piece of information is false, they continue to be influenced by that false content (Lewandowsky et al. 2012). This may be due to a combination of selective exposure and the way algorithms are organised. Selective exposure is a theory that argues that individuals will systematically seek out and adopt information that is consistent with their pre-existing beliefs while strongly avoiding anything that contradicts those beliefs ('Selective Exposure Theory'). Technology platforms take advantage of this phenomenon and build it into their algorithms, so that audiences continuously consume content they already agree with. Instagram published a transparency report on the platform's algorithm, which prioritises the order of posts on users' timelines. Instagram users' feeds are primarily sorted by their own activity: *"Posts you've liked, shared, saved or commented on help us [Instagram] understand what you might be interested in."* As a result, when users come across posts that contradict their beliefs, they ignore them. As a result, researchers and technology companies have been scrambling to find ways to counteract selective exposure bias. One such method is pre-bunking, an innovative way to pre-emptively combat disinformation.

Pre-bunking is derived from 'inoculation theory', a method borrowed from social psychology (Klepper 2022). Similar to a vaccine that creates immunity from disease, inoculation theory presents readers with a sample of potential disinformation and fights the virus by identifying falsehoods or logical fallacies. Like media literacy initiatives, when readers subsequently encounter actual disinformation, pre-bunking is designed to make readers less susceptible to believing false claims.

Pre-bunking is based on stimulating the audience's critical thinking skills and trying to build up their mental defences. Pre-bunking aims to educate the audience by identifying common manipulation techniques and logical fallacies, including appeals to emotion, personal attacks on character, and exaggerated claims. Jigsaw, a Google-led technology unit , published 'A Practical Guide to Prebunking Misinformation' (Roozenbeek et al. 2022). The report outlined three components of a successful pre-bunking message:

1. Warning: Alert users of attempts to manipulate them
2. Pre-emptive refutation: Explain the narrative/technique and how it is manipulative
3. Microdose: A weakened or practical example of misinformation that is harmless

The evidence from pre-bunking initiatives is largely positive. Several studies have shown a significant change in audience susceptibility to believing false content before and after viewing a pre-bunking message. For example, in 'Testing the Efficacy of Attitudinal Inoculation Videos

to Enhance COVID-19 Vaccine Acceptance', researchers at American University tested a series of three 30-second inoculation videos to protect against COVID-19 vaccine falsehoods. They reported in 2022 that audiences who watched the prebunking videos were more likely to recognise manipulation techniques and choose to receive the COVID-19 vaccine, and were less likely to share false information about the vaccine (Piltch-Loeb et al. 2022).

Pre-bunking is a relatively new technique, but some social media platforms have experimented with it. A 2022 Science Advances study conducted six controlled laboratory experiments in which some participants watched five inoculation videos, and both treatment and control group participants then completed a survey rating the authenticity of synthetic social media posts (mimicking similar Twitter and Facebook posts). The researchers tested participants' ability to recognise and understand the inaccuracy behind common manipulation techniques. They then placed two of the pre-bunking messages as adverts on YouTube, reaching 22,632 people. The study found that the treatment group performed better, reporting that *"the proportion of correct answers to all six headlines combined is significantly higher in the treatment condition compared to the control condition."* The participants who watched the inoculation videos on YouTube were found to have *"significantly higher discernment than the control group for technique recognition, trustworthiness, and sharing."* (Roozenbeek et al. 2022) Building on this research, over the past two years, YouTube and Jigsaw have released a series of videos on YouTube called the 'Info Intervention' series, which contain pre-bunking messages. These videos tackle a range of pieces of popular misinformation – from vaccine scepticism to Ukrainian refugees. They also include videos explaining common logical fallacies (such as 'Don't let fear-mongering manipulate you'). Many believe these videos could be the next step in tackling false content on YouTube.

Despite the promising results of research studies and examples such as the YouTube advertising programme, pre-bunking can be difficult to implement successfully. For pre-bunking messages to be effective, the audience needs to actively listen and consume the information. This is more difficult, as individuals may not be interested in watching a pre-bunking message – it may be too long, or not engaging enough. For example, in the YouTube pre-bunking video research study, only about a fifth of people watched the pre-bunking message to the end. The percentage of these people actively listening and absorbing the information may be even lower. In addition, to be truly effective, pre-bunking messages need to be distributed across multiple social platforms, requiring collaboration between pre-bunking researchers and private sector technology companies. Overall, pre-bunking certainly offers encouraging possibilities for combating misinformation and disinformation. However, the difficulties associated with this method encounters cannot be ignored, and different approaches need to be considered as well.

2.2 Anti-disinformation solutions: De-platforming and demonetising

Some of the most effective ways to tackle misinformation and disinformation revolve around reducing the likelihood that users will encounter false information in their social media feeds. Private sector and civil society organisations have so far concluded that deplatforming and demonetising could be useful in reducing false content online. Deplatforming, which involves either removing the content or reducing its visibility, reduces the amount of attention that false

content could receive. Demonetising pages or accounts removes the financial incentive for misinformation and disinformation agents to spread such content (O'Connor 2017).

Deplatforming online content can be achieved by a number of different methods, depending on where the content is located. Some websites have had their domain names revoked. The Daily Stormer, a neo-Nazi online newspaper notorious for publishing content denying the Holocaust, was removed from both Google search engines and domain name provider GoDaddy (Mettler and Selk 2017). The domain name provider revoked access to The Daily Stormer after the online newspaper published articles mocking the death of a counter-protester at the Unite the Right rally in Charlottesville, Virginia, where hundreds of white supremacists protested the removal of a statue of Robert E. Lee (Denton 2017). On social media, agents who spread misinformation and disinformation have been downgraded in the platform's algorithms, where the account is less visible, or banned. In 2020, Facebook and X (formerly Twitter), had banned the account of David Icke (a former English footballer) for spreading COVID-19 falsehoods (Coronavirus: David Icke Facebook page removed over COVID-19 conspiracy theories, 2020).

This makes it harder for users to find 'The Daily Stormer' or David Icke's controversial views. The neo-Nazi newspaper has been kicked off many internet service platforms, including Cloud-Flare, the global cloud services provider. It is currently available on the Dark Web, a part of the internet that is inaccessible via standard search engines and requires a special browser (Finkle 2021). Similarly, David Icke's account is still banned from Facebook.

However, deplatforming is not always possible. Deplatforming relies on the willingness of technology companies to set standards for deplatforming and then adhere to those standards. David Icke was reinstated, along with many other previously banned users, after Elon Musk took over leadership of X (Extremists and Conspiracy Theorists Reemerge on Twitter, 2023). Although he has not been reinstated on Facebook, videos and quotes containing David Icke's falsehoods are easily accessible there, as his fans and imitators continue to spread his ideas. In fact, a study analysing the impact of deplatforming COVID-19 disinformation found that in the seven days after Icke's account was removed in 2020, his mentions on Facebook increased by 84 percent (De-platforming Covid conspiracy theorists from Facebook has limited impact in reducing their influence, research finds, 2021). These trends suggest that deplatforming may have unintended consequences and serve to increase the visibility of the deplatformed person or account. Finally, deplatforming and demonetisation require social media platforms to hold their users accountable and take appropriate action. Although GoDaddy revoked access to the Daily Stormer's domain name in 2017, it did so only after mounting public pressure from a Twitter (X) campaign that highlighted their lack of internal motivation to deal with such false content (O'Connor 2017).

Therefore, certain companies have turned to demonetisation as a means of preventing the spread of misinformation and disinformation. Demonetisation is effective when an account or user reaches a large enough audience to receive financial benefits from the social media platform, through promoting products, enabling advertisements, or collaborating with popular brands. One of the most common forms of online financial remuneration is the sale of ad-vertisement space. As advertisers are attracted to high engagement accounts, sensationalist disinformation is a prime source of profit. NewsGuard, an online rating system for information-sharing websites and news outlets, estimates that the global advertising revenue generated

by disinformation sites in 2021 was $2.6 billion (Silverman et al. 2022). While this is a small percentage of the overall global programmatic advertising industry (which NewsGuard estimates to be $155bn in 2021) monetisation is clearly a benefit enjoyed by agents of mis/disinformation (Skibinski 2021).

As a result, many believe demonetisation could put more pressure on websites to improve their requirements for accurate content. Again, tech companies (such as Google Ads) have the greatest financial and technological capacity to effectively demonetise content or websites. Private companies and research institutes have explored various ways to circumvent the reluctance of technology companies to improve their content monitoring. One example is CheckmyAds.org, which exposes ads placed on malign websites that spread misinformation and disinformation. Many advertisers do not monitor where their content has been placed online, and this non-profit aims to alert them and increase advertiser transparency online. Once alerted, advertisers are encouraged to remove their content from bad faith websites.

2.3 Anti-extremism solutions: Hash sharing

One method of identifying extremist content online is through hash sharing, a term used by the NGO Global Internet Forum to Counter Terrorism (GIFCT) Hash-Sharing Database. A 'perceptual hash' is the digital signature of a picture, video, or PDF. Similar to blockchain technology, perceptual hashes are a numerical representation of the original content that cannot be altered – much like a fingerprint. GIFCT uses perceptual hashes to identify extremist content across multiple different platforms. Extremist content, as an umbrella term, includes a variety of images and text, from symbols of terrorist groups to recorded videos of extremist actions.

Founded in 2019 by Facebook, Microsoft, YouTube, and X, GIFCT has grown significantly since its inception, accepting membership from platforms such as Amazon, Tumblr, and Pinterest. Platforms must meet requirements outlined by GIFCT, including content moderation policies that explicitly prohibit terrorism, and a commitment to regular, public data transparency, among others. The forum began as a way to share data about ISIS-affiliated groups between platforms. Following the 2019 attacks on two mosques in Christchurch, New Zealand, GIFCT expanded its focus on ISIS-related terrorist content to include other forms of extremism, including far-right extremism. Despite GIFCT's focus on global terrorist networks, extremist content hash-sharing technology also has been used to identify local extremist expression.

When extremist content is found, the platform uploads those specific hashes to the database. Perceptual hashes focus on patterns within the content, so copycat pieces of content would be flagged even if the poster changed a few identifiable pixels. Once the perpetual hash has been uploaded, other platforms in the database can determine whether the same terrorist content is circulating on their sites. The platforms will then decide individually how to deal with the content: block it, suspend the user, or take some other step to curb the extremist expression.

The self-reported statistics look very positive. For example, in 2023, Facebook claimed to have found and dealt with more than 99.1 percent of terrorist content on its platform, with the remaining 0.99 percent being user-reported content (Dangerous Organizations: Terrorism

and Organized Hate, 2023). When the Buffalo shooter carried out his attacks, GIFCT immediately activated its Content Incident Protocol. The protocol is activated when GIFCT finds videos or recordings of violent extremist attacks on any of its member platforms. GIFCT published an update to the Content Incident Protocol for the Buffalo shooting, stating that within days of the Buffalo shooting in 2022, posters on member platforms had added 870 different hashes about the shooting to its hash-sharing database, along with many iterations of the attack video itself (Incident Response: CIP Activated in Response to Shooting in Buffalo, New York, 2022). The platforms then decided for themselves how to handle the content. This data seems promising, but self-reporting can be misleading: As the aforementioned 2021 Senate Committee report noted: *"...social media companies emphasize the volume of content they remove, rather than address why their platforms allow the proliferation of harmful content*." In short, the committee report questioned how these platforms allowed so much terrorist content to be uploaded in the first place.

One of the biggest limitations of a perpetual hash database is the lack of an enforcement mechanism. Even when content is flagged on one of the member platforms, there is no requirement for it to be removed or dealt with. Clearly, solutions to these online problems are insufficient without the full cooperation and commitment of social media platforms.

In addition, hash sharing has sparked much debate about its impact on free speech and censorship, with one free speech group even describing the database as a 'content cartel' (Windwehr and York 2020). Critics argue that the platform's use of automation unnecessarily censors social media platforms, as it relies on flawed artificial intelligence. Most platforms use AI to determine which content violates their community guidelines, resulting in some content erroneously being uploaded to the hash sharing database and subsequently removed. For example, a publication from Mnemonic, an NGO dedicated to archiving human rights violations and international crimes, found that GIFCT incorrectly flagged many ordinary social media posts as terrorist material. The publication explains that videos documenting human rights crises around the world have been removed due to mistakes made by AI algorithms, including videos of rights abuses in Yemen, Syria, and Ukraine (Caught in the Net: The Impact of Extremist Speech Regulations on Human Rights Content, 2019). GIFCT also noted in a 2022 transparency report that a quarter of the Syrian Archive's YouTube videos documenting human rights violations during the Syrian war were removed by the platform in 2017, citing terrorist content (Introducing 2022 GIFCT Working Group Outputs, 2022). The NGO is right to be concerned about this: if YouTube labels their videos as extremist content and submits the hashes to GIFCT, other platforms may follow suit, resulting in the complete removal of educational videos from the internet. This is why many, including the technology initiative Tech Against Terrorism (TAT), recommend human-in-the-loop systems, where people are more involved in the process of flagging and removing content. In the report 'Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online', TAT states that *"hashing may be insensitive to the use of the same item of content in a different context – such as journalism or academic research – and so an appeals process involving human review is also necessary."* (Macdonald et al.) In order to accurately identify extremist content, initiatives such as GIFCT need to address algorithmic issues within their programmes, particularly if they result in marginalised groups being unfairly targeted.

2.4 AI and Natural Language Processing

Researchers have developed methods to counter biases in AI language detection programmes using natural language processing (NLP), a branch of AI that aims to teach computers to understand text in a similar way to humans (What Is Natural Language Processing?). NLP breaks down a phrase or sentence into individual words and analyses the meaning of each word and how they relate to each other. It then uses sentiment analysis to determine whether the phrase or sentence is trying to convey a positive, negative, or neutral attitude. NLP is also used to identify many different factors in a piece of text: the dialect, the domain and subject matter, and cultural and religious contexts. The use of NLP can help content moderators dissect terrorist content that is somehow hidden. For example, the acronym 'ELF' might be completely innocent in a non-terrorist account. However, when posted in a different context, it could be associated with the far-left extremist group Earth Liberation Front, which conducts criminal activities against companies deemed to be harmful to the environment. Without NLP to analyse the content of the rest of the string of text, these acronyms go undetected and are allowed to remain online.

AI is based on the ability to learn. Many online moderators identify extremist content by relying on generative AI, the process of creating new and original content without direct human involvement. Online moderators sift through millions of pieces of content using pre-defined definitions of terrorist material to teach their generative AI models. The material is then used to analyse videos and images for terrorist symbols and logos, to translate speech into text, or even to identify speakers in an audio recording. Many machine-learning programmes involve analysing strings of text, such as phrases or sentences, that a user types in. AI works by being fed examples of text with terrorist themes, and then learning by using this information to identify similar matches in a dataset (Countering terrorism online with Artificial Intelligence, 2021).

However, AI is still a developing field, and has been shown to have biases. One of the biggest complications is that AI cannot infer the underlying meaning of a string of text. Instead, it identifies the direct definition of each phrase and translates it verbatim. Researchers are attempting to bridge the language processing gap in artificial intelligence by developing methods to identify the context and cultural nuances associated with a string of text. Natural Language Processing could be used to identify indicators of terrorist content by understanding hidden meaning, even in misspellings, slang words and phrases, or pop culture references.

One of the limitations of using NLP in content moderation is that culture is constantly changing and adapting. The slang and jargon of a group changes often, with one researcher even arguing that slang can vary from month to month, as a result of the dynamic online landscape (The Evolution of Language: How Internet Slang Changes the Way We Speak). Phrases that may have had a deeper meaning at one point in time may later become completely irrelevant. NLP is expected to be sophisticated enough to adapt just as quickly, which may not be entirely feasible.

In addition, NLP has mostly been used and trained for only English language processing, and is therefore currently unreliable in other languages. This is particularly challenging in a heterogeneous society where languages other than English are becoming increasingly relevant.

Facebook is using AI to identify terrorist content by using a model, XLM, that incorporates NLP for different languages. The goal of XLM is to be *"trained in one language and then used with other languages without additional training data."* (XLM-R: State-of-the-art cross-lingual understanding through self-supervision, 2021) Facebook has touted its success with XLM, claiming that it has led it to set new accuracy benchmarks for certain language translations, including German to English and Romanian to English. However, leaked documents from Facebook in 2021 revealed that 77 percent of Arabic content flagged as terrorist content had been mislabelled by the model – revealing an alarming bias in the types of languages XLM prioritises (Scott 2021).

Overall, the use of automated programmes to identify extremism requires their creators to recognise the biases of the material they feed into the AI. One of the key principles threatening the usefulness of AI is 'garbage in, garbage out.' The quality and accuracy of the input into an AI system can have a significant impact on the content that is produced. If incorrect or biased data is used in the input or training data, the AI model will learn the wrong lessons. This leads to AI systems that, when completed, would further exacerbate these biases and produce inaccurate results. In the case of social media platforms and hash sharing databases, careful attention to detail is needed to ensure that no group's content is unfairly flagged.

2.5 Identifying AI deepfakes: Adobe and Content Authenticity Initiative

Another growing issue in the evolving AI landscape is deepfakes. Deepfakes are another form of disinformation technology, an audio-visual method of deception in which auditory or visual content is altered to create a fraudulent narrative (Barney). Many existing examples of deepfakes are harmless, including humorous videos of celebrities, or creative artistic expression, but the potential for abuse is great. Home Security Heroes, a research organisation focused on online safety, has found that there were 550 percent more deepfake videos online in 2023 than in 2019, identifying more than 95,000 pieces of manipulated audio or visual content (Balobanov 2023). As AI becomes more sophisticated, the need to control the threat of manipulated content will increase.

Some deepfakes can be extremely malicious. In February 2024, Democrats in New Hampshire received a phone call purporting to be from President Joe Biden, urging them not to vote in the February presidential primary. The automated voice of the president told listeners that *"voting this Tuesday will only enable Republicans in their quest to elect Donald Trump again."*During a press conference, New Hampshire Attorney General John Fornella *"described the calls as the clearest and possibly first known attempt to use AI to interfere with an election in the U.S."* (Ramer and Swenson 2024). Without immediate action by investigators before voting began, this disinformation effort could have had disastrous consequences, manipulating voters into involuntary disenfranchisement.

Deepfake technology could also have dangerous societal implications if left unchecked. A recent social media trend has seen the proliferation of deepfake intimate photo applications that create non-consensual explicit photos of people. Some of these applications are advertised on social media platforms such as Instagram. These applications are even heavily popularised

by minors, as a school in Beverly Hills, California was investigated for non-consensual intimate photos of other children (Tenbarge 2024). Accordingly, strong measures need to be taken to control deepfakes.

Adobe has launched its Content Authenticity Initiative (CAI), which aims to *"promote transparency around the use of AI."* The tool allows users to integrate cryptographic techniques and embed metadata directly into content, including types of edits, ownership information, and metadata. The Content Authenticity Initiative supports any type of content (audio, video, documents), but has been used specifically for audio-visual media as a method to combat disinformation.

Similar to hash sharing, the CAI assigns asset hashes as signatures for pieces of content. The hash is encrypted with the original content's data, so any tampering would result in a mismatch between the original content's hash and the modified content. Any edits or modifications are recorded in the content's metadata, with descriptions of who made the edits, what tools were used, and when the edits were made. The purpose of this cryptographic technique is to allow viewers to authenticate the origin and identity of the content, with the aim that this information will be visible whenever the content is published online. The information will be displayed on partner websites under a 'Content Credentials' Icon, which would both identify and provide details of images created or edited using AI. CAI is partnered with many other major technology companies, including Photoshop and Shutterstock, as well as well-known camera companies such as Leica and Sony, both of which have committed to developing cameras with integrated CAI technology (Schneider 2024). Adobe executives have stated their intention to involve social media platforms in CAI, but despite some interest, no platforms have officially joined the initiative. The creation of a digital nutrition label through a content credential icon on social media could help curb disinformation efforts, harmful and dangerous deepfakes, and false audio aimed at destabilising voters.

2.6 Media literacy: The Trust Project

As a result of insufficient action by legislators and technology companies to adequately control the spread of false content, more attention needs to be paid to building resilience among internet users. Although pre-bunking has emerged as an innovative method of promoting media literacy, traditional methods are also effective.

Media literacy *"builds an understanding of the role of media in society as well as an essential skill of inquiry and self-expression necessary for citizens of a democracy."* (Media Literacy: A Definition and More) The process builds a sense of critical thinking in relation to media and news. In an era of manipulated information that can spread rapidly through the online landscape, media literacy helps internet users to develop scepticism about the complex environment of online information exchange. Individuals should be able to distinguish between fact, fiction, opinion, and propaganda. In addition, media literacy teaches individuals to understand the deeper ethical and societal implications of online media consumption, especially as the digital world shapes perceptions and attitudes and drives social change.

With this in mind, some states are attempting to implement media literacy in public schools by integrating media skills into core curricula. In 2023, New Jersey became the first US state to require media literacy for K-12 students. (Burney 2023) The new law aims to help students determine the authenticity of information to combat the prevalence of misinformation online. Governor Phil Murphy (D-NJ) signed the bill, stating that *"our democracy remains under sustained attack through the proliferation of disinformation,"* and that *"it is our responsibility to ensure our nation's future leaders are equipped with the tools necessary to identify fact from fiction."* Other states, including California, have taken note and begun the process of implementing media literacy in their schools, as well (McDonald 2024).

While state legislatures have taken decades to implement such measures, some organisations have taken steps to improve online media literacy. One example is The Trust Project, led by award-winning journalist Sally Lehrmann. The organisation created the 'Trust Indicators', the first global transparency standards to help audiences understand *"who and what is behind a news story."* The Trust Indicators are 8 standard disclosures that provide information about the journalist's experience, type of work, credentials, methods, local sourcing, diverse voices, actionable feedback, and general best practises of the source. The Trust Indicators can be found on The Trust Project's own website page, which also lists the news organisations that have partnered with the project. The news partners' own digital pages also include descriptions of each of the Trust Indicators, demonstrating their commitment to fighting disinformation and unethical journalism. Partnering with The Trust Initiative or similar media literacy organisations demonstrates a greater commitment to transparency and accountability, which is crucial in today's digital landscape.

3 Conclusion

Just as technological innovations are often neutral by design, but take on malign or benign characteristics depending on how they are used by users, the same philosophy applies to potential solutions to combat disinformation. Used responsibly, the tools highlighted in the preceding pages offer voters the best opportunity to become well-informed citizens, armed with credible content that will enable them to make the best decision on election day. As a non-partisan institution, we make no value judgement on what 'best' means in this case, other than to emphasise that users have access to facts. What they ultimately do with that information is up to them, their conscience, and their vote.

With just a few months to go until election day in November, we need to be realistic about the prospects for these tools to undo years of damage in creating an information ecosystem designed to entertain, generate revenue, and sometimes mislead. The notion of information as a public good was an illusion that has long since disappeared. But there is a new generation of digital natives, not just in the United States, but around the world, who will determine the shape and fate of democracy in the coming decades. How they view the world will be determined by the information they receive and their ability to separate fact from fiction, myth from reality. Each of the tools outlined above represents an important step towards reversing the deterioration of the information space, which has had dramatic consequences for social cohesion, trust in democratic institutions, and the belief that every vote counts. Much remains to be done, but we have to start from somewhere. If nothing else, this report aims to be a starting point for future voters to restore a political system that needs care, maintenance, and periodic reinvention.

References

Abernathy, Penelope Muse. *The State of Local News | Local News Initiative*. Local News Initiative. 16 November 2023. https://localnewsinitiative.northwestern.edu/projects/state-of-local-news/2023/report/#ethnic-communities. Last accessed 30 April 2024.

Adgate, Brad. *Local News Losing Billions In Revenue Each Year From Digital Media Giants*. Forbes. 17 May 2021. www.forbes.com/sites/bradadgate/2021/05/17/local-news-losing-billions-in-revenue-each-year-from-digital-media/?sh=6a1f970b474f. Last accessed 30 April 2024.

Ali, K., et al. *Viruses Going Viral: Impact of Fear-Arousing Sensationalist Social Media Messages on User Engagement*. Semantic Scholar. 3 May 2019. www.semanticscholar.org/paper/Viruses-Going-Viral%3A-Impact-of-Fear-Arousing-Social-Ali-Zain-ul-abdin/f1f1b11fdd743d12f9e4f7f2db7fe8e20176a2e0. Last accessed 30 April 2024.

Balobanov, Kirill. *2023 State Of Deepfakes: Realities, Threats, And Impact*. Home Security Heroes. 2023. www.homesecurityheroes.com/state-of-deepfakes/#overview-of-current-state. Last accessed 30 April 2024.

Barney, Nick. *What is deepfake AI? A definition from TechTarget*. TechTarget. www.techtarget.com/whatis/definition/deepfake. Last accessed 30 April 2024.

Blackburn, Menendez Lead Effort to Protect Free Speech Online. Blackburn Senate. 22 May 2020. www.blackburn.senate.gov/2020/5/blackburn-menendez-lead-effort-protect-free-speech-online. Last accessed 30 April 2024.

Buffalo Shooting Online Platform Investigative Report. New York State Attorney General. 18 October 2022. https://ag.ny.gov/sites/default/files/buffaloshooting-onlineplatformsreport.pdf. Accessed 30 April 2024.

Burney, Melanie. *New Jersey Becomes First State to Require Media Literacy for K-12*. Government Technology. 3 November 2023. www.govtech.com/education/k-12/new-jersey-becomes-first-state-to-require-media-literacy-for-k-12. Last accessed 30 April 2024.

Cassidy, Klobuchar, Kennedy Introduce Bill to Save Local Journalism | U.S. Senator Bill Cassidy. Bill Cassidy. 11 April 2023. www.cassidy.senate.gov/newsroom/press-releases/cassidy-klobuchar-kennedy-introduce-bill-to-save-local-journalism/. Last accessed 30 April 2024.

Caught in the Net: The Impact of Extremist Speech Regulations on Human Rights Content. Mnemonic. May 2019. https://mnemonic.org/en/content-moderation/impact-extremist-human-rights. Last accessed 30 April 2024.

Coronavirus: David Icke Facebook page removed over COVID-19 conspiracy theories. Sky News. 3 May 2020. https://news.sky.com/story/coronavirus-david-icke-facebook-page-removed-over-covid-19-conspiracy-theories-11981821. Last accessed 30 April 2024.

COUNTERING TERRORISM ONLINE WITH ARTIFICIAL INTELLIGENCE. the United Nations. 2021. www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf. Last accessed 30 April 2024.

Dangerous Organizations: Terrorism and Organized Hate. Facebook. 2023. https://transparency.fb.com/reports/community-standards-enforcement/dangerous-organizations/facebook/. Last accessed 30 April 2024.

Denton, Jack. *DID GOOGLE AND GODADDY SET A DANGEROUS PRECEDENT BY DROPPING A NEO-NAZI WEBSITE?* Pacific Standard. 17 August 2017. https://psmag.com/social-justice/did-google-and-godaddy-set-a-dangerous-precedent-by-dropping-a-neo-nazi-website. Last accessed 30 April 2024.

De-platforming Covid conspiracy theorists from Facebook has limited impact in reducing their influence, research finds. Cardiff University. 15 November 2021. www.cardiff.ac.uk/news/view/2584232-de-platforming-covid-conspiracy-theorists-from-facebook-has-limited-impact-in-reducing-their-influence,-research-finds. Last accessed 30 April 2024.

The Evolution of Language: How Internet Slang Changes the Way We Speak. Southern Tide Media. www.southerntidemedia.com/the-evolution-of-language-how-internet-slang-changes-the-way-we-speak/. Last accessed 30 April 2024.

Extremists and Conspiracy Theorists Reemerge on Twitter. ADL. 9 February 2023. www.adl.org/resources/blog/extremists-and-conspiracy-theorists-reemerge-twitter. Last accessed 30 April 2024.

False News. Transparency Center. 2024. https://transparency.fb.com/policies/community-standards/false-news/. Last accessed 30 April 2024.

Finkle, Jim. *Neo-Nazi group moves to 'Dark Web' after website goes down.* Reuters. 2021. www.reuters.com/article/idUSKCN1AV1HY/. Last accessed 30 April 2024.

Fioroni, Sarah. *Local News Most Trusted in Keeping Americans Informed About Their Communities.* Knight Foundation. 19 May 2022. https://knightfoundation.org/articles/local-news-most-trusted-in-keeping-americans-informed-about-their-communities/. Accessed 30 April 2024.

Global War on Terror. George W. Bush Library. www.georgewbushlibrary.gov/research/topic-guides/global-war-terror. Last accessed 30 April 2024.

Haugen, Frances. *Untitled.* Senate Committee on Homeland Security and Governmental Affairs. 15 September 2022. www.hsgac.senate.gov/wp-content/uploads/imo/media/doc/221116_HSGACMajorityReport_DomesticTerrorism&SocialMedia.pdf. Last accessed 30 April 2024.

Horwitz, Jeff. *Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt*. WSJ, 13 September 2021, www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353. Last accessed 30 April 2024.

Incident Response: CIP Activated in Response to Shooting in Buffalo, New York. GIFCT. 14 May 2022. https://gifct.org/2022/05/14/cip-activated-buffalo-new-york-shooting/. Last accessed 30 April 2024.

Introducing 2022 GIFCT Working Group Outputs. GIFCT. 2022. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-LF-Privacy-1.1.pdf. Last accessed 30 April 2024.

Jones, Seth G. *The Rise of Far-Right Extremism in the United States*. CSIS. 7 November 2018. https://www.csis.org/analysis/rise-far-right-extremism-united-states. Last accessed 30 April 2024.

Klepper, David. *'Pre-bunking' shows promise in fight against misinformation*. AP News. 24 August 2022, https://apnews.com/article/technology-misinformation-eastern-europe-902f436e-3a6507e8b2a223e09a22e969. Last accessed 30 April 2024.

Lambert, Hannah R. Fox News. 9 March 2023. www.foxnews.com/politics/domestic-extremism-bill-criminalize-free-speech-create-ministry-truth-advocacy-group-warns. Accessed 30 April 2024.

Lewandowsky, Stephan, et al. *Misinformation and Its Correction: Continued Influence and Successful Debiasing*. Psychological Science in the Public Interest. vol. 13, no. 3. 2012. https://journals.sagepub.com/doi/full/10.1177/1529100612451018.

Macdonald, Stuart, et al. *Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online*. Tech Against Terrorism, https://static1.squarespace.com/static/63e0c75f41ff767f07530a6f/t/65a5349648ce18428d686126/1705325720146/TATE+-+AI+REPORT+FINAL+%281%29.pdf. Last accessed 30 April 2024.

McDonald, John. *Jeff Share weighs in on California's new media literacy requirement*. UCLA School of Education & Information Studies. 22 January 2024. https://seis.ucla.edu/news/jeff-share-weighs-in-on-californias-new-media-literacy. Last accessed 30 April 2024.

Media Literacy: A Definition and More. Center for Media Literacy. 5 March 2024. www.medialit.org/media-literacy-definition-and-more. Accessed 30 April 2024.

Mettler, Katie, and Avi Selk. *GoDaddy – then Google – ban neo-Nazi site Daily Stormer for disparaging Charlottesville victim*. Washington Post. 14 August 2017. www.washingtonpost.com/news/morning-mix/wp/2017/08/14/godaddy-bans-neo-nazi-site-daily-stormer-for-disparaging-woman-killed-at-charlottesville-rally/. Last accessed 30 April 2024.

Missouri Attorney General Andrew Bailey Obtains Court Order Blocking the Biden Administration from Violating First Amendment | Attorney General Office of Missouri. Missouri Attorney General. 2022. https://ago.mo.gov/missouri-attorney-general-andrew-bailey-obtains-court-order-blocking-the-biden-administration-from-violating-first-amendment/. Last accessed 30 April 2024.

Mosseri, Adam. *Instagram Ranking Explained.* Instagram. 31 May 2023. https://about.instagram.com/blog/announcements/instagram-ranking-explained/. Last accessed 30 April 2024.

Newton, Casey, and Nick Barclay. *The deep inequalities of Facebook's secretive cross-check moderation program.* The Verge. 7 December 2022. www.theverge.com/2022/12/7/23498030/facebook-moderation-scandal-xcheck-cross-check. Last accessed 30 April 2024.

O'Connor, Clare. *GoDaddy, Google Ban Neo-Nazi Site Daily Stormer Following Outrage Over Charlottesville Story.* Forbes. 14 August 2017. www.forbes.com/sites/clareoconnor/2017/08/14/godaddy-bans-neo-nazi-site-daily-stormer-following-outrage-over-charlottesville-story/?sh=3e43b748659a. Last accessed 30 April 2024.

O'Hara, Kelly, and Natalie Campbell. *What is Section 230 and Why Should I Care About It?* Internet Society. www.internetsociety.org/blog/2023/02/what-is-section-230-and-why-should-i-care-about-it/?gad_source=1&gclid=Cj0KCQjwqpSwBhClARIsADlZ_Tm-QOn73DsGQ_zlem5kT3SXaNDy8zdpgyyIJYb_ffcjI5xfkcJFRMQaAoJKEALw_wcB. Last accessed 30 April 2024.

Piltch-Loeb, Rachael, et al. *Testing the Efficacy of Attitudinal Inoculation Videos to Enhance COVID-19 Vaccine Acceptance: Quasi-Experimental Intervention Trial.* JMIR Public Health and Surveillance, 2022.

Place, Nathan. *Fake news got more engagement than real news on Facebook in 2020, study says.* The Independent. 5 September 2021. www.independent.co.uk/news/world/americas/fake-news-facebook-misinformation-study-b1914650.html. Last accessed 30 April 2024.

Ramer, Holly, and Ali Swenson. *Fake Biden robocall investigation targets 2 Texas companies.* AP News. 7 February 2024. https://apnews.com/article/biden-robocalls-artificial-intelligence-new-hampshire-texas-a8665277d43d05380d2c7594edf27617. Last accessed 30 April 2024.

Rivera, Jessica M., et al. *Evaluating interest in off-label use of disinfectants for COVID-19*. NCBI. 28 September 2020. www.ncbi.nlm.nih.gov/pmc/articles/PMC7521872/. Last accessed 30 April 2024.

Rosenblatt, Kalhan, et al. *Senate hearing highlights: Lawmakers grill CEOs from TikTok, X and Meta about online child safety.* NBC News. 31 January 2024. www.nbcnews.com/tech/live-blog/senate-hearing-online-child-safety-big-tech-live-updates-rcna136235. Last accessed 30 April 2024.

Roozenbeek, Jon, et al. *A Practical Guide to Prebunking Misinformation.* Prebunking. 2022. https://interventions.withgoogle.com/static/pdf/A_Practical_Guide_to_Prebunking_Misinformation.pdf. Last accessed 30 April 2024.

Roozenbeek, Jon, et al. *Psychological inoculation improves resilience against misinformation on social media.* Science Advances. vol. 8, no. 34. 2022. Science Advances, www.science.org/doi/10.1126/sciadv.abo6254.

Schneider, Jaron. *Cameras, Content Authenticity, and the Evolving Fight Against AI Images.* Peta-Pixel. 2 January 2024. https://petapixel.com/2024/01/02/cameras-content-authenticity-and-the-evolving-fight-against-ai-images/. Last accessed 30 April 2024.

Scott, Mark. *Facebook did little to moderate posts in the world's most violent countries.* Politico. 25 October 2021. www.politico.com/news/2021/10/25/facebook-moderate-posts-violent-countries-517050. Last accessed 30 April 2024.

Selective Exposure Theory. The Decision Lab. https://thedecisionlab.com/reference-guide/psychology/selective-exposure-theory. Last accessed 30 April 2024.

Senate Judiciary Committee Advances Bipartisan Bill to Preserve Strong, Independent Journalism and News Organizations | United States Senate Committee on the Judiciary. Senate Judiciary Committee. 15 June 2023. www.judiciary.senate.gov/press/releases/senate-judiciary-committee-advances-bipartisan-bill-to-preserve-strong-independent-journalism-and-news-organizations. Last accessed 30 April 2024.

Silverman, Craig, et al. *How Google's Ad Business Funds Disinformation – ProPublica.* ProPublica. 29 October 2022. www.propublica.org/article/google-alphabet-ads-fund-disinformation-covid-elections. Last accessed 30 April 2024.

Singh, Kanishka. *Anti-Muslim incidents jump in US amid Israel-Gaza war.* Reuters. 28 January 2024. www.reuters.com/world/us/anti-muslim-incidents-jump-us-amid-israel-gaza-war-2024-01-29/. Last accessed 30 April 2024.

Skibinski, Matt. *Special Report: Top brands are sending $2.6 billion to misinformation websites each year.* NewsGuard. 2021. www.newsguardtech.com/special-reports/brands-send-billions-to-misinformation-websites-newsguard-comscore-report/. Last accessed 30 April 2024.

The State of Local News 2023 | Local News Initiative. Local News Initiative. 2023. https://local-newsinitiative.northwestern.edu/projects/state-of-local-news/2023/. Last accessed 30 April 2024.

Stocking, Galen, et al. *2. Alternative social media sites frequently identify as free speech advocates.* Pew Research Center. 6 October 2022. www.pewresearch.org/journalism/2022/10/06/alternative-social-media-sites-frequently-identify-as-free-speech-advocates/. Last accessed 30 April 2024.

Tenbarge, Kat. *Deepfake app ads on Instagram undressed 16-year-old Jenna Ortega.* NBC News. 5 March 2024. www.nbcnews.com/tech/internet/deepfake-jenna-ortega-fake-nude-image-meta-ig-instagram-facebook-rcna141023. Last accessed 30 April 2024.

Tenbarge, Kat. *Police open criminal investigation into Beverly Hills AI-nude photos incident*. NBC News. 28 February 2024. www.nbcnews.com/tech/tech-news/beverly-hills-ai-nude-photos-middle-school-vista-rcna140965. Last accessed 30 April 2024.

UNITED STATES. Human Rights Watch. 11 September 2023. www.hrw.org/reports/2002/usa-hate/usa1102-04.htm. Last accessed 30 April 2024.

Vespa, Jonathan, and Lauren Medina. *Demographic Turning Points for the United States: Population Projections for 2020 to 2060*. Census Bureau. 2018. https://www.census.gov/content/dam/Census/library/publications/2020/demo/p25-1144.pdf. Last accessed 30 April 2024.

Washington State Legislature. Washington State Legislature. 2024. https://app.leg.wa.gov/billsummary?BillNumber=1333&Initiative=false&Year=2023. Last accessed 30 April 2024.

What is a filter bubble? | Definition from TechTarget. TechTarget. www.techtarget.com/whatis/definition/filter-bubble. Last accessed 30 April 2024.

What Is Natural Language Processing? IBM. www.ibm.com/topics/natural-language-processing. Last accessed 30 April 2024.

White, Edward Douglass. *Schenck v. United States :: 249 U.S. 47 (1919)*. Justia US Supreme Court Center. https://supreme.justia.com/cases/federal/us/249/47/. Last accessed 30 April 2024.

White Supremacist Propaganda Soars to All-Time High in 2022. ADL. 8 March 2023. www.adl.org/resources/report/white-supremacist-propaganda-soars-all-time-high-2022. Last accessed 30 April 2024.

Windwehr, Svea, and Jillian C. York. *One Database to Rule Them All: The Invisible Content Cartel that Undermines the Freedom of Expression Online*. Electronic Frontier Foundation. 27 August 2020. www.eff.org/deeplinks/2020/08/one-database-rule-them-all-invisible-content-cartel-undermines-freedom-1. Last accessed 30 April 2024.

XLM-R: State-of-the-art cross-lingual understanding through self-supervision. Meta AI. 7 November 2019. https://ai.meta.com/blog/-xlm-r-state-of-the-art-cross-lingual-understanding-through-self-supervision/. Last accessed 30 April 2024.

Younis, Mohamed, and Claire Evans. *The Year in Review: 2023's Most Notable Findings*. Gallup News. 26 December 2023. https://news.gallup.com/opinion/gallup/547508/year-review-2023-notable-findings.aspx. Last accessed 30 April 2024.

Publishing credits & legal notice

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
www.bertelsmann-stiftung.de

Upgrade Democracy
www.upgradedemocracy.de

Responsible for content
Bertelsmann Foundation North America (BFNA), USA

Responsible for the publication series
Cathleen Berger
Co-Lead Upgrade Democracy
cathleen.berger@bertelsmann-stiftung.de
www.upgradedemocracy.de
www.bertelsmann-stiftung.de

Charlotte Freihse
Project Manager Upgrade Democracy
charlotte.freihse@bertelsmann-stiftung.de
www.upgradedemocracy.de
www.bertelsmann-stiftung.de

Author
Shwetha Rao

Design
nach morgen

Copyeditor
Lara Wagner

Citation note
Rao, Shwetha (2024). *Countering disinformation in the United States*. Bertelsmann Stiftung, Gütersloh. DOI: 10.11586/2024070

DOI number
10.11586/2024070