GOVERNING THE MACHINE: COUNTERING AI-DRIVEN DISINFORMATION

By DR HELENA IVANOV

HJS Henry Jackson Society

CENTRE FOR RESILIENT SOCIETY

DEMOCRACY | FREEDOM | HUMAN RIGHTS

May 2026

Published in 2026 by The Henry Jackson Society

The Henry Jackson Society
Millbank Tower
21-24 Millbank
London SW1P 4QP

Registered charity no. 1140489
Tel: +44 (0)20 7340 4520

www.henryjacksonsociety.org

Title: “GOVERNING THE MACHINE: COUNTERING AI-DRIVEN DISINFORMATION”
By Dr Helena Ivanov

£9.95 where sold

Cover image: A robotic arm playing chess against a human on an electronic chessboard by Pavel Danilyuk at pexels.com (https://www.pexels.com/photo/robot-picking-a-chess-piece-8438957/).

GOVERNING THE MACHINE: COUNTERING AI-DRIVEN DISINFORMATION

By DR HELENA IVANOV

HJS
Henry
Jackson
Society

DEMOCRACY | FREEDOM | HUMAN RIGHTS

CENTRE FOR RESILIENT SOCIETY

May 2026

About the Author

Dr Helena Ivanov is an Associate Research Fellow at the Henry Jackson Society. She holds a PhD in International Relations from the London School of Economics and Political Science. Her research focuses on the relationship between propaganda and violence against civilians. In her thesis, Dr Ivanov examined the role propaganda played during the Yugoslav Wars and produced a model for studying propaganda which details the key phases, functions, discourses, and techniques of propaganda (the model itself is applicable to other contexts). Additionally, Dr Ivanov served as a Manager at the Centre for International Studies at the LSE. Prior to her PhD, Dr Ivanov completed an MPhil in Politics: Political Theory at the University of Oxford, and holds a BA in Politics from the University of Belgrade.

Acknowledgements

I thank Andrew Fox and Natalia Cote-Muñoz for reviewing this paper and providing valuable feedback that helped strengthen and improve the report. I also thank HJS's interns who have helped with the production of this report.

Contents

About Us

HJS
Henry
Jackson
Society

DEMOCRACY | FREEDOM | HUMAN RIGHTS

About The Henry Jackson Society

The Henry Jackson Society is a think-tank and policy-shaping force that fights for the principles and alliances that keep societies free, working across borders and party lines to combat extremism, advance democracy and real human rights, and make a stand in an increasingly uncertain world. The Henry Jackson Society is a company limited by guarantee registered in England and Wales under company number 07465741 and a charity registered in England and Wales under registered charity number 1140489.

For more information, please see **www.henryjacksonsociety.org**.

CENTRE FOR RESILIENT SOCIETY

About the Centre for Resilient Society

The **Centre for Resilient Society (CRS)** is a citizen-focused, international research centre within the Henry Jackson Society, which seeks to identify, diagnose and propose solutions to threats to the social resilience of liberal Western democracies.

The centre's work includes addressing the twin challenges posed by radicalisation and terrorism. The centre is unique in addressing violent and non-violent extremism. By coupling high-quality, in-depth research with targeted and impactful policy recommendations, it aims to combat the threat of radicalisation and terrorism in our society.

The centre's work also includes broader challenges of democratic resilience – including threats from both foreign interference and domestic issues. This includes the potential harm that various forms of social, cultural and political insecurity, conflict and disengagement can pose to the long-term sustainability of democracies, including the resilience of their institutions, public policy outcomes, citizens' health and wellbeing, and economic growth and prosperity. It also explores the balance between free speech and hate speech, and encourages respectful debate between those of different views, rather than cancellation. Moreover, it underscores how social and political instability can make nations vulnerable to internal and external actors seeking to deepen cleavages, undermine consensus and, ultimately, to weaken democratic functioning.

Executive Summary

This report examines the dangers that artificial intelligence poses in the context of disinformation warfare. Disinformation, identified as a top-tier risk by the World Economic Forum, continues to threaten the stability of democratic states by eroding public trust in institutions, media, and the broader information environment. The Henry Jackson Society has written extensively on this issue, and across multiple reports has demonstrated why existing approaches have proven insufficient to resolve the problem.

As the world was still grappling with disinformation, a new and revolutionary technology emerged – artificial intelligence. While AI is poised to transform societies for the better in many respects, it is equally clear that it has the potential to shift the dynamics of disinformation warfare in a far more dangerous direction. AI can generate and disseminate disinformation at scales, speeds and costs that risk turning an already serious problem into a severe one.

In this report, we examine how we arrived at this point, why existing approaches have fallen short, and how AI is likely to exacerbate the problem unless meaningful changes are introduced. To that end, we put forward five key policy recommendations.

First, we advocate for the **use of AI in fact-checking**, supported by human oversight. Second, we call for **state-funded training programmes for educators**. Third, we recommend **mandatory media literacy programmes** integrated across the UK education system. Fourth, we argue for **radical transparency** as a necessary step toward rebuilding trust in key institutions. Finally, we emphasise the need for **constant updating and sustained strategic investment**, recognising that any policy response must evolve in line with the speed of AI development.

Introduction

The rapid advance of artificial intelligence is reshaping the global information environment and introducing a new generation of challenges for governments and democratic societies. Among the most urgent is AI's capacity to enable the large-scale creation and dissemination of disinformation. Tools capable of generating highly realistic deepfakes, synthetic news content, and automated influence campaigns are dramatically lowering the cost while increasing the speed and scale at which manipulative narratives can be produced and spread. As these systems continue to evolve, distinguishing between authentic and AI-generated content is becoming increasingly difficult.

AI-generated content can take many different forms. For the purposes of this report, however, we focus on four types that we identify as particularly harmful: false but not intentionally deceptive content, commonly referred to as misinformation; intentionally deceptive content, commonly referred to as disinformation; foreign information manipulation, which is often disinformation but specifically created or disseminated by hostile foreign actors; and, finally, synthetic image-based abuse, including non-consensual intimate imagery of real people.

The deployment of AI to generate these types of content risks fundamentally transforming both the nature of information warfare and the dynamics of political communication. Malign state and non-state actors can now use AI to micro-target audiences with personalised propaganda, amplify polarising content, and erode trust in institutions, elections, and the media. As access to these technologies expands, disinformation operations are likely to become both more sophisticated and more pervasive.

Yet this challenge did not emerge in a vacuum. Western governments have already struggled to respond effectively to disinformation in the digital age. Since the rise of social media, hostile actors – most notably Russia and, increasingly, China – have invested heavily in efforts to manipulate online discourse, deepen social divisions, and undermine democratic legitimacy. Research by the Henry Jackson Society (HJS) indicates that even digitally native populations, such as UK university students, frequently struggle to distinguish credible information from manipulation. [1] Broader evidence suggests this vulnerability is widespread across Western societies. For example, according to the 2025 Reuters Digital Report, as much as "58% of respondents say they feel unsure about their ability to distinguish truth from falsehood in online news." [2] In short, societal resilience to disinformation remains significantly underdeveloped.

The rapid development of AI threatens to exacerbate these existing weaknesses. [3] Generative systems are increasingly capable of producing convincing synthetic images, audio, and video that blur the boundary between reality and fabrication. At the same time, AI dramatically reduces the resources required to conduct influence operations, enabling forms of manipulation that were once labour-intensive to be partially or fully automated. The implications extend beyond geopolitics into domestic information environments, where AI has already been misused for harmful purposes, including the creation of non-consensual pornographic content. [4]

1 Theo Zenou and Helena Ivanov, "Breaking the Echo Chamber: Enhancing Disinformation Resilience in the UK", The Henry Jackson Society, 18 November 2025, https://henryjacksonsociety.org/publications/breaking-the-echo-chamber-enhancing-disinformation-resilience-in-the-uk/.

2 "Reuters Digital Report 2025: Falling Trust and the Rise of Alternative Media Ecosystems", IFJ, 4 July 2025, https://www.ifj.org/media-centre/news/detail/article/reuters-digital-report-2025-falling-trust-and-the-rise-of-alternative-media-ecosystems.

3 Sakshee Singh and Alan Jagolinzer, "How Cognitive Manipulation and AI Will Shape Disinformation in 2026. Here's How to Build Resilience", World Economic Forum, 12 March 2026, https://www.weforum.org/stories/2026/03/how-cognitive-manipulation-and-ai-will-shape-disinformation-in-2026/.

4 Jason Wilson, "Hundreds of Nonconsensual AI Images Being Created by Grok on X, Data Shows", *The Guardian*, 8 January 2026, https://www.theguardian.com/technology/2026/jan/08/grok-x-nonconsensual-images.

At the same time, rejecting artificial intelligence outright is neither feasible nor desirable. The technology offers significant benefits, including in the national security domain, where advanced predictive tools have already demonstrated their potential to enhance strategic foresight and decision-making (see for example Rhombus Power's role in predicting the Russian invasion of Ukraine). [5] The challenge, therefore, is not whether we should continue to develop AI, but how its use – particularly in the information domain – can be effectively governed.

Like every transformative technology before it, AI cannot be allowed to integrate into modern information ecosystems without meaningful oversight. Whether it ultimately strengthens or destabilises democratic societies will depend on the policy frameworks established today. An unregulated trajectory risks intensifying disinformation and enabling new forms of manipulation, while overly restrictive approaches risk restricting innovation and weakening the strategic position of democratic states.

This report addresses this governance gap by examining how AI is transforming the production and dissemination of disinformation and by offering concrete policy recommendations for the United Kingdom. Its aim is not to halt technological progress, but to shape it – ensuring that democratic societies can enjoy the benefits of artificial intelligence while mitigating its risks, strengthening institutional resilience, and safeguarding the integrity of public discourse in an increasingly synthetic information environment.

Put simply, AI will scale, personalise, and accelerate existing social-media vulnerabilities. In some cases, the government has already recognised the risk – and banned the creation and sharing of deep fake images. We applaud that decision. However, there are many far more ambiguous areas that require further actions.

It is precisely for those "grey area" cases that we are developing a set of policy recommendations to help governments address the problem. Our approach is intentionally more cautious and less restrictive. There are two primary reasons for this. First, broad or indiscriminate restrictions risk undermining freedom of expression – one of the foundational principles of democratic societies. Second, as the experience of social media regulation demonstrates, even where it is theoretically possible to define categories of harmful speech, implementation remains a significant challenge. Governments have repeatedly struggled to effectively curb the spread of disinformation in practice, raising serious questions about the enforceability of sweeping regulatory approaches.

Accordingly, in these cases, our policies acknowledge that some degree of disinformation will persist. Rather than pursuing maximalist restrictions that are difficult to implement and potentially counterproductive, we prioritise measures that limit the reach and impact of harmful content while strengthening societal resilience.

Ultimately, as AI continues to reduce the cost and increase the scale of disinformation, efforts to control supply alone are unlikely to succeed. A more sustainable approach lies in complementing targeted regulation with strategies that enhance public resilience, ensuring that democratic societies are better prepared to navigate an increasingly complex and synthetic information environment.

5 Laurie Clarke, "When Ripples Become Tidal Waves: Can We Predict When the next Global Crisis Will Hit?", *BBC News*, 23 April 2026, https://www.bbc.com/future/article/20260421-can-we-predict-the-next-crisis.

The (dis)information ecosystem

Disinformation has long been recognised as a major threat to democratic societies. As early as 2013, the World Economic Forum (WEF) identified "digital wildfires" as a significant risk in its annual Global Risks Report. The report warned that: "the global risk of *massive digital misinformation* sits at the centre of a constellation of technological and geopolitical risks ranging from *terrorism* to *cyber attacks* and the *failure of global governance*. This risk case examines how hyperconnectivity could enable 'digital wildfires' to wreak havoc in the real world." [6] Even at that stage, the WEF recognised the destabilising potential of social media platforms and their ability to rapidly disseminate information on a global scale.

To illustrate these risks, the report pointed to real-world examples where misinformation spread online had tangible consequences, including during Hurricane Sandy, as well as incidents in Mexico of "mothers needlessly keeping their children from school and shops closing due to false rumours of shootouts spreading through social networks." [7] According to the report, digital wildfires are most dangerous "in situations of high tension, when false information or inaccurately presented imagery can cause damage before it is possible to propagate accurate information. [...] The other dangerous situation is when information circulates within a bubble of likeminded people who may be resistant to attempts to correct it." [8]

Over the next decade, these potential dangers materialised with striking clarity. Disinformation has spread at unprecedented speed and scale, partly due to the mass adoption of social media platforms where accountability for false or misleading content is limited at best, and non-existent at worst. [9] According to Datareportal at the start of April 2026: "there were 5.79 billion social media 'user identities' around the world. [...] Social media user numbers continue to grow too, with 294 million new user identities starting to use social media since this time last year. [...] The latest figures indicate that 94.7 percent of the world's internet users regardless of age now use social media each month." [10] Moreover, they estimate that "the typical internet user spends more than 1 full working day each week using social media. And added together, the world spends over 15 billion hours consuming content on social platforms each day, which is the equivalent of more than 1.75 million years of human existence." [11] These figures alone demonstrate the sheer scale of social media's global reach. This, in turn, shows how in an environment where accountability is minimal and virtually anyone can disseminate whatever content they choose, disinformation is able not only to thrive, but to exert profound societal and political influence.

6 "Insight Report: Global Risks 2013, Eighth Edition", World Economic Forum, 2013: 23, https://www3.weforum.org/docs/WEF_GlobalRisks_Report_2013.pdf.

7 For further details see: "Insight Report: Global Risks 2013, Eighth Edition", World Economic Forum.

8 Ibid.

9 Zeve Sanderson and Scott Babwah Brennen, "We Failed The Misinformation Fight. Now What?", NYU's Center for Social Media, AI, and Politics, 25 August 2025, https://csmapnyu.org/impact/policy/we-failed-the-misinformation-fight-now-what.

10 "Global Social Media Statistics", DataReportal – Global Digital Insights, https://datareportal.com/social-media-users#:~:text=Here%20are%20some%20global%20social%20media%20statistics:,*%20**WhatsApp**%203%20billion%20monthly%20active%20users. Note: the authors of the report recognise that number of user identities is likely not the same as the number of actual human individuals as some individuals may have more than one account.

11 Ibid.

The Harm of Disinformation

And indeed, the political outcomes of the last few years demonstrate the power of disinformation. Moreover, they show how hostile actors can abuse disinformation to achieve their strategic aims. For instance, many have highlighted (although the exact impact remains disputed) the role of Russian-orchestrated disinformation and the impact it had on key political developments, including the Brexit referendum [12] and the election of Donald Trump in 2016. [13] More recently, disinformation proved central to the information warfare surrounding Russia's full-scale invasion of Ukraine in 2022. [14] Generally speaking, both domestic and foreign actors have exploited social media as a key vector for disseminating polarising, fake, and misleading narratives. Moreover, hostile states such as Russia and China are now utilising disinformation systemically and on a much larger scale than before, using it to influence public opinion, weaken democratic cohesion, and advance geopolitical objectives across democratic societies. [15]

Crucially, the consequences of disinformation extend well beyond politics. Its impact on public health has been both significant and, at times, very dangerous. During the Covid19 pandemic, governments worldwide struggled to achieve necessary immunisation levels in large part due to the proliferation of false claims about the safety and efficacy of vaccines. This wave of disinformation undermined trust in public health authorities and contributed to vaccine hesitancy at a critical time. According to experts who investigated vaccine uptake in the United States during the pandemic, disinformation has played a negative role. The report found "a negative relationship between misinformation and vaccination uptake rates. Online misinformation is also correlated with vaccine hesitancy rates taken from survey data." [16] Likewise, other researchers "studied the top shared health web links on Polish social media platform [sic] and found that 40% of the most frequently shared links contain fake news. [...] Among these health-related fake news, vaccine-related news has the most fallacious content." They further show that "misinformation induced a decline in intent of 6.2% in the UK and 6.4% in the USA among those who previously intended to take the vaccine." [17]

12 See for example: Marco Bastos, "Social Media 'Bots' Used to Boost Political Messages during Brexit Referendum", City St George's, University of London, 8 April 2022, https://www.citystgeorges.ac.uk/research/impact/case-studies/social-media-bots-used-to-boost-political-messages-during-brexit-referendum; and Rachel Ellehuus, "Did Russia Influence Brexit?", CSIS, 21 July 2020, https://www.csis.org/blogs/brexit-bits-bobs-and-blogs/did-russia-influence-brexit; and Naja Bentzen, "Online Disinformation and the EU's Response", European Parliament Research Service, February 2019, https://www.europarl.europa.eu/RegData/etudes/ATAG/2018/620230/EPRS_ATA(2018)620230_EN.pdf.

13 See for example: Hunt Allcott and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election", *Journal of Economic Perspectives* 31, no.2 (2017): 211–36, https://doi.org/10.1257/jep.31.2.211.

14 See for example: Tom Willaert and Marc Tuters, "From Denazification to the Golden Billion: An Inductive Analysis of the Kremlin's Weaponisation of Digital Diplomacy on Telegram", *Humanities and Social Sciences Communications* 12(1), 3 July 2025, https://doi.org/10.1057/s41599-025-05382-x; and Cameron Lai, Fujio Toriumi and Mitsuo Yoshida, "A Multilingual Analysis of pro-Russian Misinformation on Twitter during the Russian Invasion of Ukraine", *Scientific Reports* 14(1), 2 May 2024, https://doi.org/10.1038/s41598-024-60653-y; and Dominique Geissler and Stefan Feuerriegel, "Analyzing the Strategy of Propaganda Using Inverse Reinforcement Learning: Evidence from the 2022 Russian Invasion of Ukraine", *Arxiv*, 24 July 2023, https://doi.org/10.48550/arXiv.2307.12788; and Todd C. Helmus and Khrystyna Holynska, "Ukrainian Resistance to Russian Disinformation: Lessons for Future Conflict", *RAND*, 3 September 2024, https://www.rand.org/pubs/research_reports/RRA2771-1.html.

15 See for example: Niklas Swanström and Filip Borges Månsson (eds), "The Convergence of Disingormation: Examining Russia and China's Partnership in the Digital Age", *Institute for Security & Development Policy*, December 2024, https://www.isdp.eu/wp-content/uploads/2024/12/ISDP-Special-Paper-Disinformation.pdf; and Tamás Matura, "Sino-Russian Convergence in Foreign Information Manipulation and Interference: A Global Threat to the US and Its Allies", *CEPA*, 30 June 2025, https://cepa.org/comprehensive-reports/sino-russian-convergence-in-foreign-information-manipulation-and-interference/; and Dana S. LaFon, "How the U.S. Can Counter Disinformation From Russia and China", *Council on Foreign Relations*, 14 August 2024, https://www.cfr.org/articles/how-us-can-counter-disinformation-russia-and-china.

16 Francesco Pierri, et al., "Online Misinformation Is Linked to Early COVID-19 Vaccination Hesitancy and Refusal", *Scientific Reports* 12(1), 26 April 2022, https://doi.org/10.1038/s41598-022-10070-w.

17 Hanjia Lyu, Zihe Zheng and Jiebo Luo, "Misinformation versus Facts: Understanding the Influence of News Regarding COVID-19 Vaccines on Vaccine Uptake", *Health Data Science*, 2022, https://doi.org/10.34133/2022/9858292.

Finally, a major research project conducted in 2022 across different countries reveals a very grim picture – one from which it is clear that disinformation knows no borders and had a serious impact worldwide during the pandemic. Specifically, the authors of the report tell us that "based on responses gathered from over 18,400 individuals from 40 countries, [they found] a strong association between perceived believability of COVID-19 misinformation and vaccination hesitancy. [Their] study shows that only half of the online users exposed to rumours might have seen corresponding fact-checked information. Moreover, depending on the country, between 6% and 37% of individuals considered these rumours believable." [18]

Yet disinformation in this domain neither began nor ended with Covid19. Conspiracy theories pertaining to vaccines have now become mainstream by and large thanks to social media platforms. For example, "a survey by the Royal Society for Public Health found that 50% of British parents of children younger than 5 years regularly encountered negative messages about vaccination on social media." [19] The real-world effect of anti-vaccine propaganda is now increasingly visible: declining vaccination rates have contributed to renewed outbreaks of measles across parts of the Western world, [20] while Eastern and Northern Europe have recently faced a resurgence of whooping cough. [21]

Much more concerningly, evidence suggests that these trends are unlikely to reverse in the near future unless something substantial changes. HJS's previous research shows that more than 60% of UK-based students would feel reluctant to give their child the MMR jab based on the information they read online. [22] Even if only a half of our respondents were to follow through on this hesitancy, vaccination rates would fall dangerously below the threshold needed for herd immunity, creating conditions for serious outbreaks of preventable disease.

In parallel, the virality of social media has created incentives for individuals to promote misleading, fake, and often dangerous claims for personal gain. [23] Likewise, many others, unable to distinguish between truth and falsehoods, often unknowingly spread fake news. The case of Belle Gibson is particularly illustrative: she built a vast network of followers (north of 200,000 people) by falsely claiming to have cured terminal brain cancer through diet. She was able to achieve impressive commercial success – including automatic app integration in the Apple Watch, and a book deal with Penguin Books – before it was revealed that, in fact, Ms Gibson never had brain cancer. [24] By that stage, however, the damage had already been done. And it wasn't merely done by Ms Gibson, who intentionally spread lies, but also by many others who naively believed Ms Gibson and subsequently shared her posts. Although we cannot reliably assess how Gibson's claims influenced individual medical decisions, she undoubtedly represents an important part of the rise in scepticism of conventional medicine that is increasingly evident today. Moreover, the fact that Gibson was able to secure commercial deals as significant as those with Penguin Books and Apple also highlights just

[18] Karandeep Singh, et al., "Misinformation, Believability, and Vaccine Acceptance over 40 Countries: Takeaways from the Initial Phase of the COVID-19 Infodemic", *PLoS ONE* 17(2), 9 February 2022, https://doi.org/10.1371/journal.pone.0263381.

[19] Talha Burki, "Vaccine Misinformation and Social Media", *The Lancet Digital Health* 1(6), October 2019, https://doi.org/10.1016/S2589-7500(19)30136-0.

[20] Ibid.

[21] Mine Durusu Tanriover, et al., "The Ongoing Challenge of Pertussis in Eastern and Northern Europe: Recommendations from the Global Pertussis Initiative", *Infectious Diseases and Therapy* 15(5), 2 April 2026: 1175–201, https://doi.org/10.1007/s40121-026-01329-0.

[22] Zenou and Ivanov, "Breaking the Echo Chamber: Enhancing Disinformation Resilience in the UK".

[23] Marianna Spring, "How X Users Earn Thousands from US Election Misinformation and AI Images", *BBC News*, 19 October 2024, https://www.bbc.co.uk/news/articles/cx2dpj485nno.

[24] Emma Clifton, "Everything You Need to Know To Get Caught Up On Belle Gibson – The Craziest F-king Story You Will Ever Hear", *Capsule NZ*, 7 February 2025, https://capsulenz.com/be/everything-you-need-to-know-to-get-caught-up-on-belle-gibson-the-craziest-f-king-story-you-will-ever-hear/.

how little even major corporations may invest in properly verifying information when they are blinded by the prospect of profit.

Due to all of these developments, more than a decade after its initial warning, the WEF has elevated the risk of disinformation. In its 2024 Global Risks Report, misinformation and disinformation were ranked as the most severe global risk over a two-year horizon, and the fifth most severe over a ten-year period – highlighting both the scale of the challenge and the urgency of developing effective responses. [25]

Unfortunately, in the two years that followed, and despite disinformation rising further up the global risks agenda, no real improvement was made. Despite significant attention from governments, platforms, and international organisations, disinformation has remained a persistent and unresolved challenge, with few effective solutions emerging in practice. The WEF recognises this, and in their 2026 Global Risk Report, misinformation and disinformation were elevated to the fourth most severe risk over a ten-year horizon, while in the short term they have moved down to second place, surpassed only by geoeconomic confrontation. Given the escalation of global conflicts, ongoing wars, and the increasing use of tariffs and economic coercion between major powers, this shift does not reflect an improvement in the disinformation environment, but rather a worsening of other systemic risks competing for the top spot on the global agenda.

Naturally, this raises a fundamental question: how did we get here? If disinformation was identified as a systemic global risk as early as 2013, and a top risk in 2024, how is it that so few meaningful solutions have been developed? Even more concerningly, how has the problem not only persisted, but in many ways become significantly worse?

[25] "The Global Risks Report 2024, 19th Edition", World Economic Forum, 2024, https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf; and Bentzen, "Online Disinformation and the EU's Response".

Enter Social Media Platforms

A large part of the answer lies in the structure and operation of social media platforms themselves. It goes without saying that social media platforms have democratised communication in many ways. In authoritarian regimes, where citizens often have limited access to impartial media outlets, social media platforms have played a crucial role in disseminating information, exposing state-controlled narratives, and bringing people together.

However, by allowing information to spread so fast across the globe and giving individuals unprecedented access to mass audiences, social media platforms have also created an environment that is exceptionally fertile for disinformation. These platforms allow virtually anyone to post almost anything instantaneously, often with very limited oversight or accountability. The widespread use of anonymous or pseudonymous accounts further contributes to this problem, making it difficult to identify, let alone hold accountable, those responsible for creating or disseminating harmful falsehoods. [26]

Perhaps even more importantly, social media platforms are not designed to prioritise truth – they are designed to prioritise engagement. More specifically, "algorithms typically optimize users' revealed preferences, i.e. user engagement such as clicks, shares and likes." [27] In practice, this means that factual but less emotionally provocative content is often deprioritised, while posts that trigger outrage, fear, or anger are systematically amplified. Whether such content is accurate is, in many cases, secondary to whether it performs well.

This dynamic creates an ideal environment for disinformation to flourish. Disinformation is rarely just about spreading falsehoods; rather, it is often specifically designed to provoke emotional responses that make manipulation more effective. Emotional language, fearmongering, sensationalism, and polarisation are frequently central to its success.

Consider, for example, the difference between a balanced, factual post discussing immigration trends and an emotionally charged post warning of catastrophic societal collapse due to uncontrolled immigration. The latter is far more likely to provoke strong reactions, generate engagement, and therefore be further promoted by platform algorithms. If misleading or entirely false claims are embedded within that emotionally charged narrative, social media systems can effectively supercharge their dissemination. The result is a perfect storm in which falsehoods, when paired with emotional manipulation, are often rewarded with greater visibility than objective reality.

Further research confirms these worries. For example, Germano, Gómez and Sobbrio show how "boosting sharing-based engagement (e.g., likes or reshares) amplifies misinformation and ideological polarization." [28] Specifically, they argue that "since individuals with more extreme beliefs are disproportionately likely to highlight content aligned with their priors, this mechanism amplifies the visibility of ideologically extreme items. As a result, the distribution of content that users see and click on becomes more bimodal and polarized. In turn, this translates into higher overall platform engagement, as the higher visibility of more

[26] Press Room, "Dissemination of Misinformation via Anonymous Accounts on Right-Wing Channels", *Disinformation Social Media Alliance*, 25 December 2024, https://disa.org/dissemination-of-misinformation-via-anonymous-accounts-on-right-wing-channels/.

[27] Smitha Milli, et al., "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media", *PNAS Nexus* 4(3), 5 March 2025, https://doi.org/10.1093/pnasnexus/pgaf062.

[28] Fabrizio Germano, Vicenç Gómez, and Francesco Sobbrio, "Ranking for Engagement: How Social Media Algorithms Fuel Misinformation and Polarization", *Journal of Public Economics* 255 (2026): 105589, https://doi.org/10.1016/j.jpubeco.2026.105589.

polarized content increases the probability of individuals with more extreme beliefs reading something they are willing to highlight. Crucially, this amplification of the extremes reduces the average informational quality of consumed content, measured by the proximity of signals to the true state." [29] They continue to demonstrate not only how social media platforms have become fertile ground for disinformation, but also how they actively intensify polarisation while simultaneously increasing the perception of polarisation itself. As more balanced or moderate perspectives are comparatively deprioritised due to lower engagement, they are effectively crowded out by more extreme, emotionally charged content. Ultimately, this growing polarisation carries substantial political consequences.

Likewise, experts argue that: "Twitter's engagement-based ranking algorithm amplifies emotionally charged, out-group hostile content that users say makes them feel worse about their political out-group." [30] Thus, it is not simply that algorithms enhance the spread of disinformation and polarisation; they also negatively shape how individuals perceive those with whom they politically disagree, further deepening intergroup hostility and social division. [31]

Crucially, this ecosystem has also created powerful financial incentives for the spread of such content. While disinformation is often associated with hostile political actors or ideological agendas, increasingly it is also driven by something much simpler: profit. As platforms began monetising engagement and compensating creators, outrage itself became economically valuable. This has fuelled the rise of "rage baiting," where influencers and content creators produce inflammatory, divisive, or controversial content on purpose because they know it will trigger emotional responses and maximise engagement. As analyses by the BBC have noted, the monetisation structures of social media have directly contributed to spikes in such behaviour. [32]

Even more problematically, social media platforms are also designed in ways that actively create and reinforce echo chambers and filter bubbles. While the two concepts are similar in many respects, they are not identical. Specifically, "filter bubbles describe environments in which algorithmic curation immerses users in attitude-consistent information, whereas echo chambers emphasize active selection, where individuals choose to interact primarily with like-minded others." [33]

Regarding filter bubbles, "algorithms and hashtags amplify content by targeting specific audiences and promoting posts aligned with the user's interests and behaviours", [34] while often filtering out or deprioritising information that challenges those views. According to HJS's previous research, the algorithm is quite successful at creating filter bubbles, with more than 50% of our respondents stating that the content they see on their social media pages is aligned with their pre-existing beliefs. [35] More worryingly, our research also showed that people intentionally create echo chambers – with more than 70% of our respondents saying that they "purposefully follow accounts that reinforce their political or social beliefs while avoiding those that challenge them." [36] The end result is an information environment where

[29] Germano, Gómez and Sobbrio, "Ranking for Engagement: How Social Media Algorithms Fuel Misinformation and Polarization".

[30] M. Carroll, et al., "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media", *Knight First Amendment Institute*, 2024, https://knightcolumbia.org/content/engagement-user-satisfaction-and-the-amplification-of-divisive-content-on-social-media.

[31] S. Awasthi, "From Clicks to Chaos: How Social Media Algorithms Amplify Extremism", *Observer Research Foundation*, 2025, https://www.orfonline.org/expert-speak/from-clicks-to-chaos-how-social-media-algorithms-amplify-extremism.

[32] For further details see: S. Gruet and M. Lawton, "What Is Rage-Baiting and Why Is It Profitable?", *BBC News*, 2024, https://www.bbc.co.uk/news/articles/c4gp555xy5ro.

[33] S. He and Y. Fan, "Emotion as a Cross-Layer Mechanism in Filter Bubbles: A Social-Psychological Perspective", *Frontiers in Psychology* 16 (2025): 1740709, https://doi.org/10.3389/fpsyg.2025.1740709.

[34] Awasthi, "From Clicks to Chaos".

[35] Zenou and Ivanov, "Breaking the Echo Chamber: Enhancing Disinformation Resilience in the UK".

[36] Ibid.

users are increasingly surrounded by ideological reinforcement rather than genuine debate or corrective perspectives.

The Henry Jackson Society has already examined this phenomenon, but its implications for disinformation cannot be overstated. Propaganda and disinformation are significantly more effective when they align with existing biases or previously held beliefs. This is a well-established dynamic in information warfare. For example, Russian propaganda following its invasion of Ukraine proved highly effective among many domestic Russian audiences, in part because narratives portraying Ukrainians as extremists or Nazis built upon pre-existing prejudices and state-cultivated historical narratives. [37] By contrast, those same propaganda efforts were largely ineffective – and often counterproductive – among Ukrainians themselves, whose lived experiences and national identity fundamentally contradicted those claims. Rather than weakening Ukrainian resistance, such propaganda often strengthened it. [38]

When applied to social media ecosystems, this dynamic becomes particularly dangerous. Imagine an individual who believes that vaccines cause autism. In a balanced information environment, that person may encounter a mixture of perspectives – some reinforcing their beliefs, others challenging them. At the very least, exposure to competing viewpoints creates the possibility for reassessment or doubt. However, in an algorithmically driven filter bubble or echo chamber, that same individual may instead be exposed almost exclusively to accounts, communities, and content that reinforce their existing scepticism. In such an environment, there is little reason for that person to question their beliefs. On the contrary, repeated exposure to numerous sources making the same false claims can create the illusion of overwhelming evidence or consensus.

This is where the danger compounds. The sheer volume of reinforcing content can strengthen false beliefs over time, making them increasingly resistant to correction. Moreover, these echo chambers and filter bubbles rarely remain confined to a single issue. An individual drawn into anti-vaccine content may also be algorithmically pushed toward broader conspiratorial or anti-establishment narratives, including scepticism toward mainstream medicine, institutions, or science more generally. As social media systems identify patterns in user engagement, they often recommend adjacent forms of misinformation, creating an expanding web of interconnected falsehoods.

The consequences can be (and in fact are) profound. What begins as exposure to one form of disinformation can evolve into a broader worldview shaped by disinformation across multiple domains, from healthcare to politics to national security. In this way, social media echo chambers do not simply contribute to the maintenance of false beliefs – they can actively radicalise and deepen them. Combined with algorithmic amplification, low accountability, and financially incentivised outrage, these systems create an extraordinarily powerful infrastructure for the sustained dissemination of disinformation.

In effect, modern social media ecosystems have created a system in which virality consistently outweighs accuracy, emotional manipulation often outperforms facts, and the spread of polarising or misleading information is frequently rewarded both algorithmically and financially. This is a key reason why disinformation has proven so difficult to combat: it is no longer merely a political or social problem, but one that is increasingly inseparable from the technological and economic foundations of the digital information environment itself.

However, even this does not capture the full scale of the problem. Perhaps even more concerningly, social media platforms are not simply designed to distribute content – they are

[37] For further details see Roozenbeek, 2024.

[38] For further details see Roozenbeek, 2024.

deliberately structured in ways that make users psychologically dependent on continuous engagement. A growing number of psychiatrists, psychologists, and behavioural experts have analysed how activities such as doomscrolling, compulsive posting, and endlessly refreshing feeds can trigger dopamine responses similar to those associated with other addictive behaviours, including substance use or gambling. Features such as doomscrolling and constant notifications, as well as social validation mechanisms, create powerful feedback loops that encourage users to remain online for prolonged periods. [39] To put things into perspective, it is worth reiterating the point cited above: "the world spends over 15 billion hours consuming content on social platforms each day, which is the equivalent of more than 1.75 million years of human existence." [40]

This means that the danger is not solely rooted in the nature of the content itself, but also in the frequency and intensity with which individuals are exposed to it. Social media is not merely a passive information source; for many, it has evolved into an outright addiction.

The consequences are hard to ignore. People are no longer only encountering misleading or manipulative narratives occasionally – they are often exposed to them continuously, sometimes for hours each day, within systems specifically engineered to maximise retention. When this addictive consumption is combined with algorithmic amplification, emotionally provocative material, and echo chambers that reinforce pre-existing beliefs, the result is a very fertile environment for disinformation to thrive. In effect, many users are spending large portions of their lives hooked to digital platforms that repeatedly feed them content designed to provoke emotion, sustain attention, and often reinforce misinformation.

Obviously, it was only a matter of time before hostile actors began systematically exploiting this environment. Over the past decade, states such as Russia, increasingly China, and also Iran have invested immense resources into disinformation campaigns targeting Western democracies, with the explicit aim of destabilising societies, deepening internal divisions, undermining trust in democratic institutions, and advancing their own geopolitical objectives. [41] Russia, for the most part, has engaged in campaigns aiming to sow distrust and polarisation across the Western world, whereas China has primarily focused on using disinformation to improve its global image and push for their Taiwan agenda. According to experts, "China and Russia, while differing in style and intent, increasingly align in their [foreign information manipulation and interference] FIMI efforts, aiming to weaken Western democracies, erode public trust, and promote multipolarity through shared anti-Western narratives and media amplification. Both countries employ state-controlled media, cyber intrusions, content generated by artificial intelligence (AI), and social media influence campaigns [...] While evidence of direct coordination [between Russia and China] remains limited, their activities in discrediting democratic processes, especially around Taiwan and US elections, show alarming parallels. The use of local influencers amplifies their outreach and enhances their credibility with local audiences." [42]

[39] For further details see: Brianna Goldman, "Addictive Potential of Social Media, Explained", *Stanford Medicine News Center*, 2021, https://med.stanford.edu/news/insights/2021/10/addictive-potential-of-social-media-explained.html.

[40] "Global Social Media Statistics", DataReportal.

[41] Fellow, C.W.-R., Watts, C., and Fellow, N.-R., *Triad of disinformation: How Russia, Iran, & China ally in a messaging war against America, Alliance For Securing Democracy* (2020), available at: https://securingdemocracy.gmfus.org/triad-of-disinformation-how-russia-iran-china-ally-in-a-messaging-war-against-america/ (Accessed: 26 May 2026); and Andrew Whiskeyman and Michael Berger, "Axis of Disinformation: Propaganda from Iran, Russia, and China on COVID-19", *The Washington Institute for Near East Policy*, 25 February 2021, https://www.washingtoninstitute.org/policy-analysis/axis-disinformation-propaganda-iran-russia-and-china-covid-19; and Matura, "Sino-Russian Convergence in Foreign Information Manipulation and Interference: A Global Threat to the US and Its Allies"; and Daria Mosolova, "How AI Is Supercharging Russia's Online Disinformation Campaigns", *BBC News*, 27 February 2026, https://www.bbc.co.uk/news/articles/cx2r7grrdwzo.

[42] Matura, "Sino-Russian Convergence in Foreign Information Manipulation and Interference: A Global Threat to the US and Its Allies".

Unfortunately, in many respects, these efforts have been effective. While it would be overly simplistic to claim that foreign disinformation alone has determined major political outcomes, it has unquestionably contributed to increased polarisation, weakened social cohesion, and heightened distrust across democratic societies. Today, many Western democracies are demonstrably less stable, more politically fragmented, and more vulnerable to manipulation than they were prior to the social media era. And as the Center for European Policy Analysis (CEPA) shows: "While Russian and Chinese foreign information manipulation and interference (FIMI) operations may appear to affect only the soft power of the US and its allies, these actions have a fundamental impact on hard power and national security as well." [43]

But the problem of how disinformation is created, and how social media platforms amplify its dissemination, is only one part of the picture. The second part concerns the recipients themselves – many of whom, even by their own admission, remain deeply vulnerable to disinformation, lacking both the resilience and the confidence to reliably identify it when they encounter it. Our previous research at the Henry Jackson Society further illustrates just how concerning this reality is. Even younger generations – those born and raised in the digital age – have not proven resilient to disinformation. Quite the opposite: many frequently struggle to distinguish false information from credible reporting, often believe demonstrably inaccurate claims, and openly acknowledge that they lack the tools or knowledge to effectively counter disinformation when they encounter it. This suggests that digital nativity does not automatically translate into digital resilience. [44]

Perhaps even more troubling is that disinformation's greatest success is not always in convincing people to believe outright falsehoods, but in sowing confusion on such a scale that individuals no longer know what to believe at all. [45] In many cases, the objective is not persuasion but epistemic destabilisation – undermining confidence in the very possibility of objective truth. Once citizens lose trust not only in institutions but also in the broader information environment itself, democratic discourse becomes significantly more fragile.

Indeed, years of political polarisation, institutional distrust, disinformation campaigns, and the rapid spread of anti-establishment narratives have significantly weakened public confidence in mainstream media across many democratic societies. Increasingly, large portions of the public no longer view traditional media institutions as credible or trustworthy sources of information. For example, according to Gallup's research in 2025 "Americans' confidence in the mass media has edged down to a new low, with just 28% expressing a 'great deal' or 'fair amount' of trust in newspapers, television and radio to report the news fully, accurately and fairly. This is down from 31% [in 2024] and 40% [in 2020]." [46]

Moreover, this loss of trust, at least in the US, applies to people from all sides of the political spectrum. As Gallup reports: "Republicans' confidence, which hasn't risen above 21% since 2015, has dropped to single digits (8%) for the first time in the trend. Independents' trust has not reached the majority level since 2003, and the latest 27% reading matches [2024] historical low. For Democrats, the narrowest of majorities (51%) now express trust in the media, which is a repeat of the low previously seen in 2016." [47]

[43] Matura, "Sino-Russian Convergence in Foreign Information Manipulation and Interference: A Global Threat to the US and Its Allies".

[44] For further details see: Zenou and Ivanov, "Breaking the Echo Chamber: Enhancing Disinformation Resilience in the UK".

[45] For further details see the case of Southport stabbing and disinformation surrounding Brigitte Macron in: Zenou and Ivanov, "Breaking the Echo Chamber: Enhancing Disinformation Resilience in the UK".

[46] Brenan, Megan, "Trust in Media at New Low of 28% in U.S.", *Gallup.com*, 2025, https://news.gallup.com/poll/695762/trust-media-new-low.aspx (Accessed 26 May 2026).

[47] Ibid.

The United States is by no means an isolated example. According to the 2025 Reuters Digital Report, trust in the mainstream media outlets is falling across the world, and people are turning to "alternative media ecosystems." Specifically, "the report reveals [that traditional media] are struggling to connect with most of the public as public engagement continues to decline, contrary to a rising dependence on social media and video platforms." [48]

The entire dynamic is perhaps best understood as a vicious cycle. As the disinformation crisis has deepened, growing numbers of people have increasingly turned away from established media outlets in favour of social media platforms as their primary source of political information. Yet social media, with its limited safeguards and weak mechanisms of accountability, often exposes users to even greater volumes of disinformation. The viewers, often unable to distinguish between truth and falsehoods, fall victim to disinformation. This, in turn, further erodes trust in mainstream media, making individuals even more susceptible to future manipulation and falsehoods.

While mainstream media is by no means without flaws – bias, sensationalism, and editorial failures have all contributed to legitimate criticism – traditional media institutions nonetheless operate within frameworks that impose at least some degree of accountability, professional standards, legal liability, and regulatory oversight. Journalists and established outlets can, in principle, be publicly challenged, corrected, sanctioned, or held legally responsible for false reporting in ways that anonymous online actors or algorithmically amplified social media accounts often cannot.

However, as trust in these institutions has declined, millions of people have increasingly come to rely on social media platforms as their primary source of information. This transition is particularly dangerous because social media ecosystems are not fundamentally structured around truth, verification, or responsibility, but around engagement, virality, and user retention. In moving away from comparatively regulated media environments toward platforms where misinformation can spread with minimal oversight, societies have inadvertently deepened their vulnerability to manipulation.

In effect, the broader atmosphere of distrust and polarisation has helped create a vacuum in which mainstream media's declining authority has not been replaced by more reliable alternatives, but rather by highly fragmented and often far less accountable digital ecosystems. This has further accelerated the spread of disinformation, making democratic societies more susceptible not only to falsehoods, but also to confusion, radicalisation, and sustained domestic and foreign manipulation.

[48] Reuters Institute for the Study of Journalism and International Federation of Journalists, "Reuters Digital Report 2025: Falling Trust and the Rise of Alternative Media Ecosystems".

Policy Responses to Date

Thus, part of the reason why the World Economic Forum has repeatedly escalated its assessment of disinformation as a global risk lies in the structural realities of the modern digital ecosystem – particularly the way social media platforms operate and the fertile ground they create for disinformation to spread. It is important to recognise that governments, international organisations, technology companies, and civil society actors have invested immense time, resources, and political capital over the past decade attempting to address disinformation. But, as the WEF's own risk assessments demonstrate, these efforts have thus far failed to produce sufficiently effective results.

One of the primary tools social media platforms initially relied upon was fact-checking. [49] Professional fact-checkers were introduced to identify, challenge, and correct false or misleading information circulating online. In theory, this represented an important step toward greater informational accountability. In practice, however, the scale of the challenge quickly became apparent. No realistic number of human fact-checkers could match the speed, volume, and sophistication of coordinated disinformation networks, bot farms, or algorithmically amplified viral falsehoods. Disinformation could be produced and disseminated at industrial scale, while corrections were often slower, less visible, and less emotionally compelling. [50]

Moreover, fact-checking itself became politically contentious. During the Covid19 pandemic, companies such as Meta introduced aggressive fact-checking and moderation tools to combat vaccine misinformation. [51] While these measures were often implemented with public health intentions, they also triggered significant backlash. Many users came to perceive fact-checkers not as neutral arbiters of truth, but as politically biased gatekeepers or instruments of censorship. [52] In some cases, concerns regarding overreach or inconsistency were not entirely unfounded, and the pendulum may indeed have swung too far toward restrictive moderation. And once mistakes were made, it was hard to reverse their impact. For example, "Twitter's handling of the Hunter Biden laptop story and Facebook's suppression of the COVID lab leak theory (both of which the platforms later reversed course on as new evidence came to light)" [53] caused irreparable damage to the public's perception about fact-checkers. Nevertheless, despite imperfections, fact-checking represented an attempt to impose at least some corrective mechanism on an otherwise chaotic information environment.

However, political and cultural scepticism toward fact-checking intensified further following the election of Donald Trump in 2024 and particularly after Elon Musk's takeover of X. As distrust toward institutional moderation grew, many major platforms began retreating from professional fact-checking models altogether. [54] While community-based systems such as

[49] On how fact-checking was introduced see for example: Legum, Judd, "The Facts About Facebook's Fact-Checking Program", *Popular Information*, 2020, https://popular.info/p/the-facts-about-facebooks-fact-checking (Accessed 26 May 2026); and Spangler, Todd, "YouTube Launches Fact-Checking Feature in U.S. to Fight Misinformation Amid COVID-19 Crisis", *Variety*, 2020, https://variety.com/2020/digital/news/youtube-fact-checks-us-misinformation-covid-19-1234591908/. Accessed 26 May 2026; and Thorbecke, Catherine, "What to Know About Twitter's Fact-Checking Labels", *ABC News*, 27 May 2020, https://abcnews.go.com/Business/twitters-fact-checking-labels/story?id=70903715 (Accessed 26 May 2026).

[50] See for example: Stray, Jonathan and Sneider, Eve, "Meta Dropped Fact-Checking Because of Politics. But Could Its Alternative Produce Better Results?", *Tech Policy Press*, 3 February 2025, https://techpolicy.press/meta-dropped-fact-checking-because-of-politics-but-could-its-alternative-produce-better-results/ (Accessed 26 May 2026).

[51] Rosen, Guy, "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19", *Meta*, 16 April 2020, https://about.fb.com/news/2020/04/covid-19-misinfo-update/ (Accessed 26 May 2026).

[52] See for example: Stray and Sneider, "Meta Dropped Fact-Checking Because of Politics. But Could Its Alternative Produce Better Results?".

[53] Ibid.

[54] See for example: Liv McMahon, Zoe Kleinman and Courtney Subramanian, "Facebook and Instagram get rid of fact checkers", *BBC News*, 7 January 2025, https://www.bbc.com/news/articles/cly74mpy8klo.

Community Notes emerged as alternatives, they have thus far shown limited effectiveness in combating large-scale disinformation. To illustrate: "Community Notes were designed to 'democratize' moderation, but research shows they are largely ineffective in their current form. On X, only about 1 in 10 proposed notes ever becomes visible, and they are even less likely to appear on polarized topics where they are needed most. One flaw lies in the 'consensus-based' methodology. By requiring agreement from users who usually disagree, platforms have effectively allowed partisans to hold facts hostage." [55]

As previous research by the Henry Jackson Society has argued, while reforming fact-checking systems to improve neutrality and trustworthiness was long overdue, abandoning them entirely has likely represented a step backward – once again enabling disinformation to proliferate with fewer meaningful constraints. [56]

Educational initiatives have similarly faced limitations. Across Western democracies, governments and organisations have launched numerous media literacy campaigns, workshops, and public awareness programmes aimed at improving societal resilience to disinformation. While well-intentioned, these efforts have generally failed to produce transformative results. As research from organisations like the Henry Jackson Society demonstrates, many populations remain highly vulnerable to manipulation despite such interventions. [57]

Part of the challenge is structural: educational responses have often been fragmented, inconsistent, and limited in reach. Rather than being systematically embedded into national education systems, many programmes have been voluntary, localised, or targeted primarily at already engaged populations. In other words, they frequently reach those who are already somewhat aware of the problem, while failing to meaningfully inoculate broader society. According to Bronstein and Vinogradov, "educational interventions have significant limitations: chiefly, they require individuals who are motivated to seek and voluntarily engage them. This complicates outreach to populations with lower digital media literacy. [...] furthermore, even effective educational interventions published in prominent journals do not eliminate vulnerability to misinformation. [...] A final limitation of educational interventions stems from their focus on perceived accuracy of misinformation. Perceived accuracy has little impact on information sharing, likely because social media encourages individuals to focus on other factors, such as whether sharing will attract and please followers and friends." [58]

Moreover, widespread, practical, and institutionally integrated media literacy education – particularly at school and university level – has largely been absent across much of the West. Adding fuel to fire, school and university classrooms have increasingly become the hotspots for spreading disinformation. According to experts, "this generation of children and adolescents actively avoids news, especially from traditional outlets, and mostly gets political information incidentally through social media or interpersonal sources." [59] Across the education sector, teachers and professors identify disinformation as a major problem – specifically, "a RAND Corporation study found that nearly 90% of secondary school teachers felt that students' inability to evaluate credible online information was a problem." [60]

[55] Stephan Mündges, "Community Notes Alone Won't Beat Disinformation: Why Factcheckers Are Essential", *Tech Policy Press*, 3 March 2026, https://www.techpolicy.press/community-notes-alone-wont-beat-disinformation-why-factcheckers-are-essential/.

[56] For further details see: Zenou and Ivanov, "Breaking the Echo Chamber: Enhancing Disinformation Resilience in the UK".

[57] Ibid. Also see: Diana Owen, "The Challenge of Misinformation in the Civics Classroom", *Communication Education*, 29 January 2026, pp.137–143, https://www.tandfonline.com/doi/full/10.1080/00909882.2025.2573944#d1e119.

[58] Michael V. Bronstein and Sophia Vinogradov, "Education alone is insufficient to combat online medical misinformation", *EMBO Reports* 22(3), 18 February 2021, https://pmc.ncbi.nlm.nih.gov/articles/PMC7926209/.

[59] Owen, "The challenge of misinformation in the civics classroom".

[60] Ibid.

Moreover, even where there were attempts to institute some disinformation inoculation in schools, this turned out to be a very challenging task. Teachers tasked with helping students and pupils learn how to combat disinformation "face criticism from parents and politicians who feel that school is not the place for countering misinformation. [...] The American Psychological Association (APA) found that one-third of preK-12 teachers were verbally harassed by students, and 29% were threatened by the parents of a student during the COVID-19 pandemic. The APA designated violence against educators a public health problem, as 14% of teachers reported that they were victims of physical violence from students after expressing their views." [61]

Even more problematically, the teachers themselves do not necessarily feel equipped to teach pupils and students to recognise disinformation. Specifically, "less than 20% of civic educators in a 2024 CERTL study had professional development dealing with misinformation even tangentially, and only 2% had training related to AI. In fact, only 13% of teachers, compared to 38% of students, felt that they could detect and critically analyze AI-generated content." [62]

Moreover, effective disinformation resilience requires more than theoretical instruction; it demands expert-led, realistic, and continuously updated training capable of responding to evolving technological threats. Unfortunately, such systematic educational frameworks and broader institutional responses remain largely absent. The Henry Jackson Society has previously advocated for the introduction of far more comprehensive and structured programmes, and we continue to maintain that such initiatives are essential. [63] Without them, societies will remain fundamentally underprepared for the scale and sophistication of modern disinformation threats.

In short, while substantial efforts have been made to combat disinformation, many existing approaches have either been too limited, too politically divisive, too fragmented, or too slow to meaningfully address the scale of the challenge. The result is that despite over a decade of growing awareness, disinformation remains not only unresolved, but increasingly entrenched.

[61] Owen, "The challenge of misinformation in the civics classroom".

[62] Ibid.

[63] For further details see: Zenou and Ivanov, "Breaking the Echo Chamber: Enhancing Disinformation Resilience in the UK".

Enter AI

And just as societies were still trying to figure out how to address the disinformation crisis – and the profound societal, political, and public health consequences it had already created – a new technology emerged, one likely to make the problem substantially worse: artificial intelligence. The already severe and largely unresolved challenge of disinformation is now set to intensify dramatically with the rise and exponential development of generative AI.

There are several reasons for this.

First, AI has the capacity to generate fake images, videos, and audio recordings at a scale previously unimaginable for those seeking to create and disseminate disinformation. If people have already struggled to distinguish truth from falsehood in the era of social media, this problem is bound to become significantly more severe as AI-generated deepfakes become increasingly sophisticated and far better at convincingly mimicking reality. [64] Already today, "human evaluators were not much better than chance at detecting deep fake videos. The vast scale of content production has created a 'fog of information' where authenticity is increasingly difficult to discern." [65] As these technologies continue to improve at extraordinary speed, distinguishing authentic material from fabricated content is likely to become progressively difficult for the average person.

Second, AI is likely to fundamentally transform both the cost and speed of producing disinformation. The cost of generating manipulative content will decline dramatically, while the speed at which it can be created and disseminated will increase substantially. [66] Here is an indication of what is likely coming: "estimates by the European Parliamentary Research service indicate that the number of deepfake videos shared online could surge from approximately 500,000 in 2023 to 8 million by 2025. This means that the volume of fabricated videos potentially grew 16 times during this period." [67]

If these figures are anything to go by, hostile actors will be able to produce more disinformation, more quickly, and at far lower cost than ever before. [68] Consequently, many of the already insufficient countermeasures currently in place are likely to become even less effective. If human fact-checkers struggled to compete with bot farms and coordinated troll networks in the past, there is little reason to believe they will be able to effectively counter AI-powered systems capable of automating disinformation production and dissemination at industrial scale. Moreover, the costs associated with combating AI-generated disinformation are likely to exceed those required to produce it, particularly in an environment where major social media platforms are already moving away from professional fact-checking mechanisms altogether.

Third, AI systems will likely become far more capable of tailoring disinformation campaigns at unprecedented scale. Rather than relying on broad, generic messaging, hostile actors may increasingly be able to create highly tailored narratives specifically designed to exploit the fears, beliefs, biases, or grievances of different groups or even individuals. According to the WEF,

[64] For further details see also: Singh and Jagolinzer, "How Cognitive Manipulation and AI Will Shape Disinformation in 2026. Here's How to Build Resilience".

[65] Mosolova, "The AI videos supercharging Russia's online disinformation campaigns".

[66] See for example: Mosolova, "The AI videos supercharging Russia's online disinformation campaigns".

[67] Jieun Shin, "AI in the age of fake (imagined) content", *Stimson Center*, 23 February 2026, https://www.stimson.org/2026/ai-in-the-age-of-fake-imagined-content/.

[68] See for example how Iran has utilized Lego-style AI generated propaganda: Matt Shea and Laurie Kalus, "Iran war: We spoke to the man making Lego-style AI videos that went viral", *BBC News*, 11 April 2026, https://www.bbc.com/news/articles/cjd8jrd1vnyo.

"those susceptible to potential emotional manipulation can be easily identified with micro-targeting, which uses self-reported online data to reveal personality type. Once identified, targeted messaging is selected because it resonated emotionally and will likely be shared because it affirms prior beliefs, stirs up anger or resentment, or is considered humorous. [...] Evidence from the 2024-2025 electoral cycle shows how AI systems optimized content for maximum emotional impact across multiple countries." [69] In practice, this means that large-scale disinformation campaigns may simultaneously deploy thousands or millions of customised narratives, making them substantially more believable, persuasive, and difficult to counter. Add to that the emotional aspect of messaging and you are likely to get posts with high engagement levels which will be further amplified due to the algorithmic functioning of social media platforms.

Fourth, the speed of AI development itself presents an enormous challenge. The rapid pace at which new models, tools, and features are being introduced makes it extremely difficult for the general public to keep up in ways that would allow them to effectively identify AI-generated deepfakes. The consequences will be profound: people may believe what is false and dismiss what is true. This also creates major obstacles for educators seeking to build meaningful long-term resilience. Designing educational programmes, university modules, or school-level curricula capable of preparing current or future generations becomes exceptionally difficult when the technology itself evolves faster than institutional systems can reasonably adapt. Likewise, regulators face a similar problem: any legislation or governance frameworks introduced today may quickly become outdated, effective only against current-generation systems while proving inadequate against significantly more advanced models that could become publicly accessible within months.

The deployment and consequences of using AI to sow distrust in the West are already visible. Researchers argue that "in 2024, more than 80 percent of countries experienced observable instances of AI usage relevant to their electoral processes. By far, the most popular employment of AI across elections was for content creation (accounting for 90 percent of all observed cases), compared to content proliferation (24 percent), hypertargeting (three percent) and unclear uses (four percent)." [70]

Moreover, "according to a 2024 report by the cybersecurity firm Recorded Future, at least 28 countries experienced deepfake incidents targeting public figures within a single year. Forged audio and video are now regularly attributed to political candidates, eroding public trust [even further] and manipulating opinions. In the lead-up to the 2024 U.S. primaries, one documented incident in New Hampshire involved a deepfake robocall imitating President Biden that urged voters not to cast ballots." [71] Neither America nor the UK are exceptions. Instead, over the 2024-2025 election cycle period, the utilisation of AI surged across countries. The WEF lists a couple of interesting examples: "In Ireland's 2025 presidential elections, a deepfake video falsely depicted the eventual winner withdrawing his [sic] candidature, and included fake footage of national broadcasters 'confirming' the news. [...] The Netherlands likewise saw roughly 400 AI-generated synthetic images used to attack political counterparts." [72]

Foreign and hostile state and non-state actors already began exploiting AI capabilities – most notably, Russia. For example, the BBC reported a story on King's College London professor Alan Read – one of the latest victims of AI generated deepfakes. In the video, "a synthetic voice nearly identical to that of Dr Read went on a politicised tirade against French President

69 Singh and Jagolinzer, "How Cognitive Manipulation and AI Will Shape Disinformation in 2026. Here's How to Build Resilience".

70 Cameron McKay and Inga Trauthig, "Then and now: How does AI electoral interference compare in 2025?", *Centre for International Governance Innovation*, 7 August 2025, https://www.cigionline.org/articles/then-and-now-how-does-ai-electoral-interference-compare-in-2025/.

71 Shin, "AI in the age of fake (imagined) content".

72 Singh and Jagolinzer, "How Cognitive Manipulation and AI Will Shape Disinformation in 2026. Here's How to Build Resilience".

Emmanuel Macron, berating him and other Western leaders as 'aboard the Titanic which has European Union written on its hull.'" [73] According to the BBC, "the avatar of the unwitting Dr Read appeared in a new wave of Russia-linked synthetic videos that swept across social media over the past month, raising concern among security experts that the West must brace for a battle against the Kremlin's influence on the artificial intelligence front." [74]

Just how far Russia has gotten with utilising AI for disinformation is best illustrated through the case of Romania. [75] In Romania, the 2024 "presidential election results were annulled after evidence showed AI-powered interference using manipulated videos." [76] The scale of alleged foreign interference was deemed so significant by the Court that the election itself had to be cancelled and repeated. [77]

However, as is often the case in situations of this nature, the controversy did not end with the legal decision. To this day, many Romanian citizens continue to believe that the annulment was unjustified and that the democratic will of the people was undermined. Thus, even if Russian interference did not necessarily secure the precise political outcome it may have sought, it nonetheless appears to have succeeded in something just as damaging: permanently weakening trust in Romania's democratic institutions and further destabilising public confidence in the integrity of its electoral system. [78]

However, the dangers posed by AI-generated deepfakes do not stop with elections or geopolitics. Women and children have also become major targets, particularly through the creation and dissemination of non-consensual pornographic images. [79] The process itself is alarmingly simple: users can engage with increasingly accessible, user-friendly AI chatbot interfaces, provide prompts or real images, and instruct the system to manipulate that content into virtually anything imaginable, including but not limited to digitally removing clothing from real individuals without their consent.

It did not take long for abuse to emerge at scale. Moreover, the range of victims has expanded, with female politicians increasingly targeted. The most recent victim is Italian Prime Minister Giorgia Meloni. [80] Increasingly the problem is becoming systemic. According to research conducted by Trinity College Dublin, of 500 sampled posts on X that utilised AI chatbot Grok "nearly three-quarters of posts collected and analyzed [...] were requests for nonconsensual images of real women or minors with items of clothing removed or added." [81] While X has banned the use of Grok for creation of such images, it seems that the restriction does not necessarily apply to premium subscribers. Importantly, not all governments have allowed the problem to spiral unchecked. In light of the scale of the abuse, the UK government has introduced legislation criminalising the creation and sharing of such images, with offenders facing penalties of up to two years in prison. We welcome the government's decision and urge other states to follow suit.

73 Mosolova, "The AI videos supercharging Russia's online disinformation campaigns".

74 Ibid.

75 For further details see: Sarah Rainsford, "Romanian court annuls result of presidential election first round", *BBC News*, 6 December 2024, https://www.bbc.com/news/articles/cn4x2epppego.

76 McKay and Trauthig, "Then and now: How does AI electoral interference compare in 2025?".

77 Anda Iulia Solea, "Why Romania's election was annulled – and what happens next?", *The Conversation*, 16 December 2024, https://theconversation.com/why-romanias-election-was-annulled-and-what-happens-next-245779.

78 For further details see: Ovidiu Voicu, "Romania still in the dark over election annulled due to 'hostile interference'", *Balkan Insight*, 25 December 2025, https://balkaninsight.com/2025/12/25/romania-still-in-the-dark-over-election-annulled-due-to-hostile-interference/bi/.

79 Shin, "AI in the age of fake (imagined) content".

80 Gabriele Barbati, "Meloni slams AI-generated images of herself, calling deepfakes a 'dangerous tool'", *Euronews*, 6 May 2026, https://www.euronews.com/my-europe/2026/05/06/meloni-deepfake-showing-underwear-fuels-backlash-and-ai-concerns.

81 Wilson, "Hundreds of Nonconsensual AI Images Being Created by Grok on X, Data Shows".

Policy Recommendations

And there we have it – the perfect storm. The world was already struggling to contain the disinformation crisis and by and large was failing to do so. Then, just as governments, institutions, and societies were attempting to navigate an already deeply destabilising information environment, a revolutionary new technology emerged with enormous potential to make the situation significantly worse. It didn't take long before this new technology was abused by everyone: from hostile regimes who are aiming to destabilise democracies to individuals who create pornographic content of minors and women without their consent. And this is only the beginning – as these technologies evolve and people become more adept at using them, things can become much worse very quickly.

It is beyond clear that AI is here to stay – and in many respects, for good reason. The potential advances AI offers, from medicine to scientific research to national security, are likely to be transformative. AI has already demonstrated immense value across numerous sectors, and halting its progress outright would be neither realistic nor strategically wise. Even if Western democracies hypothetically wished to pause AI development until they better understood how to mitigate its disinformation risks, that option is no longer practically available. Hostile actors, most notably China, are investing heavily in AI development, and this is a strategic race the West cannot afford to lose. To unilaterally slow innovation would risk ceding technological dominance to geopolitical competitors, repeating strategic mistakes seen in other sectors, such as overdependence on China in critical supply chains.

However, doing nothing is equally dangerous.

Russia, China, Iran, and other hostile actors are already actively working to weaponise AI in order to intensify disinformation warfare against the West. The evidence suggests that these efforts are becoming more effective. Failure to act will almost certainly further weaken democratic resilience, social cohesion, and institutional trust – outcomes that hostile actors would be more than happy to see happen. Moreover, the threat is not confined solely to foreign adversaries. Domestic bad actors also exploit the freedoms inherent in democratic systems, whether through politically motivated disinformation or the creation of deeply harmful material such as non-consensual pornographic imagery targeting women and children. Finally, many are inadvertently participating in the problem – as many individuals click on that share button genuinely believing that they are sharing something which is true, even though it is fake.

Finding the right policy balance will be exceptionally difficult. To also come up with a set of policies that are likely to be effective will be even more difficult. Overly aggressive restrictions risk undermining democratic freedoms and innovation, while weak responses risk allowing the problem to spiral further out of control. Moreover, even if someone were to halt AI's evolution, the tools that already exist and are widely available can easily wreak havoc. Research already demonstrates that relatively simple interventions – such as merely labelling deepfakes – are insufficient on their own, as every day a new tool that can remove AI watermarks is becoming available. Even when users are aware content may be manipulated, they may still emotionally or cognitively rely upon it. According to Clark and Lewandowsky, "transparency is insufficient to entirely negate the influence of deepfake videos", and their research shows that participants continued to rely "on the content of a deepfake video, even when they had been explicitly warned beforehand that it was fake. [...] this result was observed even with participants who indicated that they believed the warning and knew the video to be fake." [82]

[82] Simon Clark and Stephan Lewandowsky, "The continued influence of AI-generated deepfake videos despite transparency warnings", *Communications Psychology* 4, Article number: 13 (2026), https://www.nature.com/articles/s44271-025-00381-9?fromPaywallRec=false.

Moreover, labels can often be removed, circumvented, or ignored. Thus, while transparency measures are necessary, they are far from sufficient.

Accordingly, we propose a broader policy framework composed of multiple complementary recommendations. Crucially, this framework must remain dynamic. AI is evolving at extraordinary speed, and any policy that is static risks becoming rapidly obsolete. Constant review, adaptation, and technological updating must remain central to any effective governance model. Importantly, we commend the government for recognising the severity of the AI challenge, at least in some areas, and we welcome the recent decision to outright ban the creation and sharing of nonconsensual deepfakes. We also urge the government to continue working closely with social media platforms and AI companies to ensure that appropriate safeguards are put in place. However, as this chapter has illustrated throughout, addressing that issue alone is simply not enough. An effective response to this crisis requires a far more sophisticated framework – one capable of addressing the full spectrum of harms and vulnerabilities that AI-driven disinformation is likely to create.

At the same time, we must recognise that AI has not created the disinformation crisis; rather, AI is intensifying vulnerabilities that were already deeply embedded within the social media ecosystem. As such, our previous recommendations [83] regarding structural reform of social media platforms remain critically important. Without fundamental changes to how these platforms amplify content, prioritise engagement, and reward virality over truth, no AI-specific policy will fully resolve the broader problem.

Utilising AI for Fact-Checking

Just as AI has opened new avenues for the creation and dissemination of disinformation, it can also be a very powerful tool to counter it. AI systems can process and analyse content at scales impossible for humans, allowing for much faster identification, flagging, and proposals for countering harmful falsehoods. To illustrate, AI-powered systems can monitor vast quantities of online posts, identify suspicious patterns, identify disinformation campaigns, and propose adequate and effective responses much faster than any human fact-checker. As the European Commission's Joint Research Centre shows: "Large language models (LLMs) can analyse patterns in the dissemination of messages and narratives to spot signs of coordinated intent. This can indicate the evidence of manipulation. [...] As clustering is multilingual, stories and narratives can be identified across languages and countries, revealing how disinformation campaigns are constructed and disseminated." [84] Thus, not only can we use AI to counter disinformation, but it can also be exceptionally helpful at telling us how disinformation operates.

Obviously, AI cannot and should not do this alone. AI models can be biased, spread misinformation, and hallucinate. Thus, we argue in favour of a hybrid model in which AI is used to identify, process, and recommend; but final verification and implementation remain under human control.

Of course, we are fully aware that such systems can be exploited and circumvented. As AI becomes more embedded in the fact-checking ecosystem, hostile actors are likely to adapt, including by producing disinformation specifically designed to evade AI-based detection.

For this reason, we argue that any AI model used for these purposes must be continuously updated, stress-tested, and refined, so that evasion becomes increasingly difficult. This is

[83] For further details see: Zenou and Ivanov, "Breaking the Echo Chamber: Enhancing Disinformation Resilience in the UK".

[84] "The JRC explains: AI: friend or foe of disinformation?", European Commission Joint Research Centre, 24 September 2025, https://joint-research-centre.ec.europa.eu/jrc-explains/ai-friend-or-foe-disinformation_en.

also why the continued role of human fact-checkers remains essential. Human oversight is necessary not only to assess whether disinformation continues to spread, but also to monitor how effectively AI systems identify and counter it, and to detect any patterns of successful or systemic circumvention.

State-Funded Training Programmes for Educators

It is now beyond clear that disinformation, social media manipulation, and AI-generated falsehoods have firmly entered school and university classrooms, yet much of the education sector remains exceptionally underprepared to address them. Research, including previous work by the Henry Jackson Society, consistently demonstrates that even digitally native students are often deeply vulnerable to disinformation, lacking the necessary skills to effectively identify or combat it.

We therefore advocate for a far more systematic and structured government-led educational response. Specifically, we propose that every academic year group – from primary school through to university – should have at least one specifically designated teacher, lecturer, or professor responsible for delivering mandatory instruction focused on artificial intelligence, disinformation, social media algorithms, and digital resilience. In practice, this would mean one trained educator for first-year students, another for second-year students, and so forth throughout each stage of education.

These designated educators would be tasked with teaching students how AI systems function, how generative AI can be used to create and disseminate disinformation, how social media platforms and their algorithms amplify harmful content, how disinformation spreads, and what practical steps individuals can take to protect themselves and others – including but not limited to recognising disinformation and AI-generated content, reporting such content, and engaging individuals who disseminate such content in a safe and hopefully effective way.

However, such programmes will only be effective if the educators themselves are properly equipped. We therefore strongly advocate for substantial state funding to support specialised rolling training seminars for these educators, delivered in collaboration with leading AI companies, social media firms, technology experts, journalists, and practitioners. From companies such as Meta to Anthropic and others, those at the forefront of these technologies should play a direct role in preparing educators to teach future generations.

Given the speed at which AI technologies are developing, this training must not be static. It must operate on a continuous rolling basis, ensuring educators remain consistently updated as technologies evolve.

Mandatory Media Literacy Programmes [85]

Once sufficient institutional capacity and educator expertise have been established, we advocate for mandatory media literacy and AI resilience programmes across the entirety of the UK education system.

These programmes should not be optional, fragmented, or reactive. Rather, they should form a core and compulsory part of modern education. Every pupil or student, at every educational level, should receive one dedicated weekly class specifically focused on recognising disinformation, understanding social media ecosystems, identifying AI-generated manipulation, analysing algorithmic amplification, and developing practical resilience strategies.

[85] For further details see: Singh and Jagolinzer, "How Cognitive Manipulation and AI Will Shape Disinformation in 2026. Here's How to Build Resilience", especially the Finland approach to making children resilient to disinformation.

As noted above, this would be delivered by one designated trained educator per year group, ensuring consistent, age-appropriate instruction throughout pupils' and students' educational journey.

Importantly, these programmes should move beyond outdated or purely theoretical models. They should be interactive, practical, and regularly updated, incorporating real-world case studies, expert-led workshops, technology industry specialists, journalists, national security experts, and practitioners with direct experience combating disinformation. As much as possible they should also include practice sessions, so that pupils and students get real-life experience of recognising disinformation when they encounter it and combatting it in a safe and effective way.

The objective is not simply to teach students to passively identify fake news, but to actively understand how modern information systems operate, how manipulation occurs, and how democratic societies can build resilience against it.

Without this kind of deeply embedded, systematic educational reform, Western democracies risk continuing to produce generations that are technologically immersed but practically defenceless in the face of increasingly sophisticated AI-powered disinformation. Likewise, not making them mandatory would also be wrong – as naturally only those interested would apply, leaving everybody else exposed to disinformation.

Radical Transparency

The current crisis of trust stems, in large part, from the fact that governments, major institutions, technology platforms, and corporations have not always operated transparently. Those who create and disseminate disinformation have understood this credibility gap extremely well and have abused it to their own benefit. In many cases, when censorship or withholding of information later came to light, the damage done to public trust was irreparable.

For example, the aforementioned decisions by major platforms to suppress discussions – i.e. the possible lab leak theory of Covid19 – before reversing those positions as more evidence emerged, significantly damaged public trust in fact-checkers and institutions.

Thus, we fully advocate for full transparency regarding disinformation and policies used to combat it, including but not limited to government anti-disinformation programmes, greater transparency requirements for AI developers and social media platforms, mandatory labelling of AI-generated content, mandatory disclosure when AI tools were used in fact-checking processes, and clear public accountability structures. If Western societies wish to rebuild trust, they must start with transparency.

However, mere transparency is not enough. The obvious criticism, especially from the right, will be that these recommendations amount to state truth-management. That criticism should not simply be dismissed. In fact, it should be taken seriously precisely because any policy response to disinformation that weakens freedom of speech will ultimately fail both morally and politically.

For that reason, any anti-disinformation framework must come with serious institutional constraints. Governments should not be in the business of deciding ordinary political truth, nor should they be empowered to police legitimate domestic political disagreement. A clear difference must be established between speech that amounts to national security threats and legitimate political disagreements where differing opinions must be heard and respected.

To make that separation meaningful, anti-disinformation measures should be subject to independent oversight, published criteria, clear audit trails, regular transparency reports,

parliamentary scrutiny, and meaningful routes for appeal. Decisions should be documented, explainable, and open to challenge. The public must be able to see who acted, under what authority, according to which criteria, and with what limits. Thus, to combat disinformation governments must be transparent but also allow freedom of expression to thrive even when opinions voiced may be in stark opposition to what the government is arguing for.

Constant Updating and Strategic Investment

Finally, democratic societies must recognise that combating AI-driven disinformation is not a one-time policy challenge – it is an ongoing strategic necessity. Hostile actors are investing enormous resources into continuously improving their manipulation capabilities, and democratic responses must match or exceed that level of commitment. This means sustained financial investment, rolling policy reviews, international cooperation, and continuous adaptation.

The West has previously underestimated strategic threats – whether in defence spending, technological dependency, supply chains, or critical resources – and paid a very high price for doing so. This time, however, there is still an opportunity to act. Unlike with social media, where governments failed to appreciate the possible consequences and harms early enough, with AI we still have a chance to be proactive. We now know how dangerous disinformation is, how much it damages democracy, and we can understand the possible consequences of an uncontrolled rollout of AI in the (already broken) information ecosystems. Thus, this may be one of the final major opportunities democratic societies have to prepare and respond before the problem escalates.

Subsequently, the West can no longer afford to rely on semi-regular conferences, sporadic policy papers and reviews, or isolated and inefficient regulatory interventions. Combating AI-generated disinformation must be a constantly evolving strategic priority. To put it simply, the fact that disinformation has remained so persistently high on the World Economic Forum Global Risks Report is a clear indication that our approaches thus far have been fundamentally insufficient. We have not solved the problem. Rather, under our very eyes, it has continued to escalate. Now, with the rapid advancement of artificial intelligence, the potential for that escalation is theoretically limitless. The time has come to start treating disinformation as the top-tier strategic threat it has clearly become.

Thus, the policies that we propose above should only be seen as a starting point. We strongly advocate for continuous policy updating, large-scale strategic investment to tackle the issue, international cooperation among key stakeholders as well as cross-industry cooperation between key stakeholders, and adaptive governance.

As the Western world increases defence spending, it is worth remembering that before wars are fought on battlefields, they are almost always fought in information spaces first. By investing seriously in combating AI-powered disinformation now, democratic societies may not only protect themselves domestically, but potentially help prevent future conflicts before they escalate.

The time to act is now.

Title: “GOVERNING THE MACHINE: COUNTERING AI-DRIVEN DISINFORMATION”
By Dr Helena Ivanov

The Henry Jackson Society
Millbank Tower, 21-24 Millbank
London SW1P 4QP, UK

www.henryjacksonsociety.org

HJS
Henry
Jackson
Society

CENTRE FOR RESILIENT SOCIETY

DEMOCRACY | FREEDOM | HUMAN RIGHTS

May 2026