BRILL
NIJHOFF

SECURITY AND HUMAN RIGHTS 29 (2018) 55-81

SECURITY AND
HUMAN RIGHTS
brill.com/shrs

Artificial Intelligence and Disinformation

How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered

*Katarina Kertysova**
George F. Kennan Fellow, Kennan Institute, Woodrow Wilson Center
katarina.kertysova@gmail.com

Abstract

This article explores the challenges and opportunities presented by advances in artificial intelligence (AI) in the context of information operations. The article first examines the ways in which AI can be used to counter disinformation online. It then dives into some of the limitations of AI solutions and threats associated with AI techniques, namely user profiling, micro-targeting, and deep fakes. Finally, the paper reviews a number of solutions that could help address the spread of AI-powered disinformation and improve the online environment. The article recognises that in the fight against disinformation, there is no single fix. The next wave of disinformation calls first and foremost for societal resilience.

Keywords

disinformation – artificial intelligence – deep fakes – algorithms – profiling – micro-targeting – automated fact-checking – social media

1 Introduction

In recent years, Western democracies have been grappling with a mix of cyber-attacks, information operations, political and social subversion, exploitation of

* This report was written with research and editorial support of Eline Chivot, Senior Policy Analyst at the Center for Data Innovation. Opinions expressed in the article are solely those of the author.

existing tensions within their societies, and malign financial influence. Information operations, which constitute the focus of this paper, have been deployed by foreign actors with the objective to manipulate public opinion formation, degrade public trust in media and institutions, discredit political leadership, deepen societal divides as well as to influence citizens' voting decisions.

These challenges are playing out against the backdrop of growing digital economy, which came hand in hand with the emergence and accelerated adoption of new technologies, such as the Internet of Things (IoT), robotics and artificial intelligence (AI), 5G, or augmented and virtual reality (AR/VR). Although online disinformation is not a new phenomenon, rapid advances in information technologies – particularly the use of AI – have altered the ways in which information (and disinformation) can be produced and disseminated.[1]

Despite their numerous benefits, AI-powered systems raise a host of ethical questions and pose new risks for human rights and democratic political processes across the OSCE area. Concerns raised by the expert community include lack of algorithmic fairness (leading to discriminatory practices such as racial and gender biases), content personalisation resulting in partial information blindness ("filter bubble"), the infringement of user privacy, potential user manipulation, or video and audio manipulation without the consent of the individual.[2]

The 2016 US presidential election showed evidence of the effect digital transformation is having on democracy and political life. The use of algorithms, automation, and AI boosted the efficiency and the scope of the disinformation campaigns and related cyber activities, impacting opinion formation and voting decisions of American citizens.[3] As the role of AI in technology that powers our daily lives grows, algorithms will hold increasing sway, enabling malign actors to infiltrate government and corporate networks in order to steal information, compromise individual privacy, and distort elections without much of a trace.[4]

1 Naja Bentzen, "Computational Propaganda Techniques" (European Parliamentary Research Service (EPRS), October 2018).

2 Valerie Frissen, Gerhard Lakemeyer, and Georgios Petropoulos, "Ethics and Artificial Intelligence," Bruegel, December 21, 2018, https://bruegel.org/2018/12/ethics-and-artificial-intelligence/.

3 Philip N. Howard, Samuel Woolley, and Ryan Calo, "Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration," *Journal of Information Technology & Politics* 15, no. 2 (April 3, 2018): 81–93.

4 Jamie Fly, Laura Rosenberger, and David Salvo, "Policy Blueprint for Countering Authoritarian Interference in Democracies" (The German Marshall Fund of the United States (GMF), 2018).

This paper recognises AI applications as double-edged. While AI provides a powerful, scalable, and cost-efficient solution to prevent the distortion of information online – through the automated detection and removal of false content – it also comes with its own set of limitations and unintended consequences. This paper first examines the ways in which AI can be used to counter disinformation online. It then dives into some of the limitations of AI solutions and threats associated with AI techniques. Finally, the paper reviews a number of solutions and future developments of AI that could be envisaged to address the threat of AI-powered disinformation.

In line with the report issued by the European Commission's High Level Expert Group on Fake News and Online Disinformation, this paper defines disinformation as "false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit."[5] This paper distinguishes disinformation from misinformation – which refers to unintentionally misleading or inaccurate information[6] – and from hate speech.

Although there is no generally accepted definition of AI, the term can be understood as the ability of a system to perform tasks characteristic of human intelligence, such as learning and decision-making.[7] Machine learning (ML) can be generally defined as the usage of algorithms and large datasets to train computer systems to recognise patterns that had not previously been defined, and the ability of these systems to learn from data and discern valuable information without being programmed explicitly to do so.[8] By AI, this paper refers to ML techniques that are advancing towards AI, such as audio-visual analysis programmes that are algorithmically trained to recognise and moderate dubious content and accounts to assist human judgment.[9]

5 "A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation" (Brussels: European Commission, April 30, 2018).

6 Ibid.

7 Sophie-Charlotte Fischer, "Artificial Intelligence: China's High-Tech Ambitions," CSS Analyses in Security Policy (ETH Zurich, February 8, 2018), https://css.ethz.ch/en/center/CSS-news/2018/02/artificial-intelligence-chinas-high-tech-ambitions.html; "ITIF Technology Explainer: What Is Artificial Intelligence?" (Information Technology and Innovation Foundation (ITIF), September 4, 2018), https://itif.org/publications/2018/09/04/itif-technology-explainer-what-artificial-intelligence.

8 Fischer, "Artificial Intelligence: China's High-Tech Ambitions"; Louk Faesen et al., "Understanding the Strategic and Technical Significance of Technology for Security: Implications of AI and Machine Learning for Cybersecurity" (The Hague Security Delta (HSD), 2019).

9 Chris Marsden and Trisha Meyer, "Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism" (European Parliamentary Research Service (EPRS), March 2019).

2 AI Disinformation Solutions

There is a considerable number of initiatives aimed at countering disinformation worldwide. According to the latest figures published by the Duke Reporters' Lab, there are 194 fact-checking projects active in more than 60 countries.[10] The number of fact-checking initiatives has quadrupled in the past five years (from 44 active initiatives recorded in 2014).[11] To date, fact-checking has been mostly based on manual human intervention to verify the veracity of information. As the volume of disinformation continues to grow, manual fact-checking is increasingly judged ineffective and inefficient to evaluate every piece of information that appears online.

The first proposals to automate online fact-checking appeared a decade ago. Trump's election increased interest in the research of AI-assisted fact-checking. The last few years have seen a wave of additional funding being earmarked for automated fact-checking (AFC) initiatives that would help practitioners identify, verify, and correct social media content. To name but a few, in 2016 London-based fact-checking charity Full Fact began developing AFC tools with a €50,000 grant from Google.[12] In 2017, the charity secured an additional $500,000 (over €447,000) in funding from the Omidyar Network and the Open Society Foundations.[13] Argentinian nonprofit fact-checking organisation Chequeado and the Duke Reporters' Lab have built similar tools that scan media transcripts and identify fact-checkable claims.[14] So far, mainly independent, nonprofit fact-checking organisations have spearheaded the development and implementation of AFC.[15]

10 "Database of Global Fact-Checking Sites," Duke Reporters' Lab, n.d., https://reporterslab.org/fact-checking/.

11 Bill Adair, "Duke Study Finds Fact-Checking Growing Around the World," Duke Reporters' Lab, April 4, 2014, https://reporterslab.org/duke-study-finds-fact-checking-growing-around-the-world/.

12 Jasper Jackson, "Fake News Clampdown: Google Gives €150,000 to Fact-Checking Projects," The Guardian, November 17, 2016, https://www.theguardian.com/media/2016/nov/17/fake-news-google-funding-fact-checking-us-election.

13 "Full Fact Awarded $500,000 to Build Automated Factchecking Tools," Full Fact, June 29, 2017, https://fullfact.org/blog/2017/jun/awarded-500000-omidyar-network-open-society-foundations-automated-factchecking/.

14 Daniel Funke, "These Fact-Checkers Won $2 Million to Implement AI in Their Newsrooms," Poynter, May 10, 2019, https://www.poynter.org/fact-checking/2019/these-fact-checkers-won-2-million-to-implement-ai-in-their-newsrooms/.

15 Lucas Graves, "Understanding the Promise and Limits of Automated Fact-Checking" (The Reuters Institute for the Study of Journalism at the University of Oxford, February 2018).

Driven by computing power rather than human behavior, at first glance, AI appears to provide an impartial countermeasure against disinformation. As the accuracy and performance of AI systems continue to improve, there are growing expectations that machines can succeed where humans have failed – namely, in overcoming personal biases in decision making. As the following section shows, AI systems come with their own set of limitations and challenges.

2.1 *Algorithmic Detection of Disinformation*

In the context of information operations, AI solutions have been particularly effective in detecting and removing illegal,[16] dubious, and undesirable content online. AI techniques have also been successful in screening for and identifying fake bot accounts – techniques known as bot-spotting and bot-labelling.[17] By labelling accounts identified as bots, social media firms are enabling users to better understand the content they are engaging with and judge its veracity for themselves.[18] As regards their accuracy, however, detection algorithms need to be further developed in order to be comparable to the e-mail spam filter technology.

Google, Facebook, Twitter, and other Internet services providers rely on machine-learning algorithms to stamp out trolls, spot and remove fake bot accounts, and to proactively identify sensitive content. According to Facebook, 99.5 percent of terrorist-related removals, 98.5 percent of fake accounts, 96 percent of adult nudity and sexual activity, and 86 percent of graphic violence-related removals are detected by AI tools – not users – many of which are trained with data from its human moderation team.[19] Facebook is now moving to use similar technologies to detect false stories as well as to spot duplicates of stories that have already been debunked.[20]

Closely related to machine learning and AI is pattern recognition, which makes it possible to identify harmful online behavior. Taking a cue from articles

16 Illegal content can range from terrorist content, child sexual abuse material, incitement to hatred and violence, copyright material, and counterfeit products.

17 Eric Rosenbach and Katherine Mansted, "Can Democracy Survive in the Information Age?" (Belfer Center for Science and International Affairs, Harvard Kennedy School, October 2018), https://www.belfercenter.org/publication/can-democracy-survive-information-age.

18 Rosenbach and Mansted, 20.

19 Quoted in Marsden and Meyer, "Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism," 17.

20 Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement," Facebook, November 15, 2018, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

flagged as inaccurate by users and fact checkers in the past, AI can be leveraged to find (patterns of) words that can identify false stories.[21]

For the time being, fully automated fact-checking remains a distant goal. Social media platforms continue to rely on a combination of AI for the most repetitive work and human review for the more nuanced cases (also known as hybrid models and filters).[22] In 2018, Facebook employed 7,500 human moderators to review user content.[23] In addition, the company announced its intention to establish an independent content oversight body by the end of 2019, which will consist of external members rather than employees, and which will examine some of Facebook's most controversial content moderation decisions.[24]

2.2 *Limitations of AI Solutions*

There are several limitations to the application of automated techniques to detect and counter disinformation. The first significant shortcoming is the risk of over-blocking lawful and accurate content – the "overinclusiveness" feature of AI. The technology is still under development and AI models are still prone to false negatives/positives – i.e., identifying content and bot accounts as fake when they are not. False positives can negatively impact freedom of expression and lead to censorship of legitimate and reliable content that is machine-labelled incorrectly as disinformation.[25]

This is due to the fact that automated technologies remain limited in their ability to assess the accuracy of individual statements.[26] Current AI systems can only identify simple declarative statements, and miss implied claims or claims embedded in complex sentences, which humans recognise easily.[27] The same goes for expressions where contextual or cultural cues are necessary. AI

21 Louk Faesen et al., "Understanding the Strategic and Technical Significance of Technology for Security: Implications of AI and Machine Learning for Cybersecurity" (The Hague Security Delta (HSD), 2019).

22 Zuckerberg, "A Blueprint for Content Governance and Enforcement."

23 Christine Lagorio-Chafkin, "Facebook's 7,500 Moderators Protect You From the Internet's Most Horrifying Content. But Who's Protecting Them?," Inc., September 26, 2018, https://www.inc.com/christine-lagorio/facebook-content-moderator-lawsuit.html.

24 Brent Harris, "Global Feedback and Input on the Facebook Oversight Board for Content Decisions," Facebook, July 27, 2019, https://newsroom.fb.com/news/2019/06/global-feedback-on-oversight-board/.

25 Marsden and Meyer, "Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism," 17.

26 Marsden and Meyer, 2.

27 Graves, "Understanding the Promise and Limits of Automated Fact-Checking," 3.

systems have yet to master basic human concepts like sarcasm and irony, and cannot address more nuanced forms of disinformation.[28] Linguistic barriers and country-specific cultural and political environments further add to this challenge.

In addition, some automated algorithms run the risk of replicating and even automating human biases and personality traits, producing outcomes that are less favorable to individuals within a particular group.[29] As observers note, "however objective we may intend our technology to be, it is ultimately influenced by the people who build it and the data that feeds it."[30] Bias in algorithms can emanate from the values and priorities of those who design and train them – the programmers – or from flawed, incomplete or unrepresentative training data.[31] In computer science, the aphorism "Garbage in, garbage out" suggests that regardless of how accurate a program's logic may be, the results will be incorrect if the input is invalid.[32]

If left unchecked, observers warn, bias in algorithms may lead to decisions which can have a collective, disparate impact on certain groups of people.[33] For instance, algorithms trained on historical datasets have shown to replicate social biases, notably those against women, which then influence computer-made decisions ranging from recruitment for jobs to mortgages.[34] Whether AI can truly be freed from human error and ego is a contested topic within computer science itself.[35]

28 James Vincent, "AI Won't Relieve the Misery of Facebook's Human Moderators," The Verge, February 27, 2019, https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms.

29 Nicol Turner Lee, Paul Resnick, and Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," Brookings, May 22, 2019, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

30 Rumman Chowdhury and Narendra Mulani, "Auditing Algorithms for Bias," Harvard Business Review, October 24, 2018, https://hbr.org/2018/10/auditing-algorithms-for-bias.

31 Turner Lee, Resnick, and Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms."

32 "GIGO," TechTerms, n.d., https://techterms.com/definition/gigo.

33 Turner Lee, Resnick, and Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms."

34 Bhardwaj Gitika, "Women and the Fourth Industrial Revolution," Chatham House, June 25, 2019, https://www.chathamhouse.org/expert/comment/women-and-fourth-industrial-revolution.

35 H. Akin Ünver, "Computational Diplomacy," Cyber Governance & Digital Democracy (Centre for Economics and Foreign Policy Studies (Edam), November 2017).

Design choices can also have inherent flaws. WhatsApp provides a good illustration of the extent to which architecture and design choices may impact polarisation and misinformation. On this platform, messages are end-to-end encrypted and thus, by design, beyond the reach of content moderators. In countries like India, WhatsApp is not only a major channel for political campaigning but also a channel for false reporting and hate speech that is known to have fuelled mob-related violence and killings.[36] Because forwarded messages contain no information about the original source, there is an unclear division between official messaging and unauthorised spread of lies which, in turn, allows perpetrators to plausibly deny their involvement.[37] While encryption is a security feature, privacy of conversations is a design choice which, in essence, strips platforms of all responsibility for the content of their networks.[38]

The complexity and opacity constitute another limitation of AI systems.[39] Machine learning includes neural networks and deep neural networks, systems which are inherently "black box solutions" and whose evolution based on self-teaching goes beyond the understanding of the developers who build them.[40] The complex logic of automated decision-making makes algorithms more accurate but it is also what makes it difficult to explain how they generated a particular recommendation. A number of companies, particularly in the Silicon Valley, and the US Defense Research Agency (DARPA)'s Explainable AI Program are developing technologies and framework systems that could eventually provide verifiability, greater accountability, and transparency of machine learning.[41] For now, however, the production of explainable systems remains academic and this technology cannot be widely exploited yet.

36 Michael Safi, "'WhatsApp Murders': India Struggles to Combat Crimes Linked to Messaging Service," The Guardian, July 3, 2018, https://www.theguardian.com/world/2018/jul/03/whatsapp-murders-india-struggles-to-combat-crimes-linked-to-messaging-service.

37 John Harris, "Is India the Frontline in Big Tech's Assault on Democracy?," The Guardian, May 13, 2019, https://www.theguardian.com/commentisfree/2019/may/13/big-tech-whatsapp-democracy-india.

38 Ibid.

39 Curt Levey and Ryan Hagemann, "Algorithms With Minds of Their Own: How Do We Ensure That Artificial Intelligence Is Accountable?," Wall Street Journal (WSJ), November 12, 2017, https://www.wsj.com/articles/algorithms-with-minds-of-their-own-1510521093.

40 AJ Abdallat, "Explainable AI: Why We Need To Open The Black Box," Forbes, February 22, 2019, https://www.forbes.com/sites/forbestechcouncil/2019/02/22/explainable-ai-why-we-need-to-open-the-black-box/#5bfe81df1717.

41 David Gunning, "Explainable Artificial Intelligence (XAI): Program Update November 2017" (DARPA, November 2017), https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

It is important to note that both automated and human verification mechanisms have their limitations and unintended consequences. Human moderators often work in highly stressful conditions, under tight schedules, and can struggle to cope with traumatic images and videos. Without adequate training and support to deal with the disturbing content, moderators can develop PTSD-like symptoms, which can affect their ability to perform their day-to-day activities.[42] Second, human review is not only costly, it is also prone to error and ambiguous results, particularly when someone's background, personal ethos, or even mood on any given day might influence content analysis. "This job is not for everyone," Facebook acknowledged in 2018, detailing the hiring and training processes and how the company provides access to mental health resources in addition to paying attention to the environment where reviewers work.[43]

Lastly, AI solutions raise important questions about who is best placed to determine what content is legal or illegal, desirable or undesirable. Should the judgment about the truthfulness and urgent removal of online content lie with public entities (whether or not institutionally linked to governments), judicial authorities, or online platforms?

3 Threats Associated with AI Techniques

While advances in machine learning technologies will unarguably benefit those who defend against malign information operations online, they are also likely to allow adversaries to magnify the scale and effectiveness of their operations in the short term.[44] As Eric Rosenbach and Katherine Mansted from the Belfer Center for Science and International Affairs of Harvard Kennedy School argue, "breakthroughs are likely to spread quickly and widely, equipping both state and non-state adversaries with a technological edge."[45] While non-state actors, such as the Islamic State (ISIS), have been effective in using disinformation for recruitment purposes and will likely utilise all possible means to pursue

42 Casey Newton, "The Trauma Floor: The Secret Lives of Facebook Moderators in America," The Verge, February 25, 2019, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

43 Sasha Lekach, "'The Cleaners' Shows the Terrors Human Content Moderators Face at Work," Mashable, November 13, 2018, https://mashable.com/article/the-cleaners-content-moderators-facebook-twitter-google/.

44 Rosenbach and Mansted, "Can Democracy Survive in the Information Age?," 12.

45 Ibid., 14.

their terrorist activities, they lack the resources to scale up their operations.[46] In contrast, state actors such as Russia and China invest considerable resources in new technologies.[47] China, in particular, aims to dominate other economies in AI. The proliferation of this technology among authoritarian states constitutes a long-term risk to democratic principles.

Disruptive technologies are already finding their application in the political sphere, including for the purposes of information manipulation. Four threats stand out in particular: (1) user profiling and segmentation; (2) hyper-personalised targeting; (3) deep fakes; and (4) humans finding themselves "out of the loop" of AI systems.[48]

3.1 *User Profiling and Micro-Targeting*

With advances in machine learning, adversaries will increasingly be able to identify individuals' unique characteristics, beliefs, needs, and vulnerabilities. They will then be able to deliver highly-personalised content, and thereby target with maximum effectiveness those who are most vulnerable to influence.[49]

In the context of elections, it is important to draw a distinction between *demographic* and *psychometric* profiling. While demographic profiling is informational and segments voters based on age, education, employment, or country of residence, psychometric profiling is behavioral and enables personality-based voter segmentation.[50] Two individuals with the same demographic profile (for example two white, employed, middle-aged, single women) can have markedly different personalities and opinions. Content tailored to different personality types – whether they are introverted, extroverted, or argumentative – is more likely to evoke the desired response.[51]

In the run up to the 2016 US presidential election, presidential candidate Hillary Clinton used demographic segmentation techniques to identify groups

46 Alina Polyakova and Spencer Phipps Boyer, "The Future of Political Warfare: Russia, the West, and the Coming Age of Global Digital Competition," The New Geopolitics of Europe and Russia (The Brookings Institution, March 2018), https://www.brookings.edu/research/the-future-of-political-warfare-russia-the-west-and-the-coming-age-of-global-digital-competition/.

47 Ibid.

48 See Rosenbach and Mansted, "Can Democracy Survive in the Information Age?"

49 Ibid.

50 Michael Wade, "Psychographics: The Behavioural Analysis That Helped Cambridge Analytica Know Voters' Minds," The Conversation, March 21, 2018, https://theconversation.com/psychographics-the-behavioural-analysis-that-helped-cambridge-analytica-know-voters-minds-93675.

51 Ibid.

of voters. In addition to demographics, Cambridge Analytica – an advertising company contracted to the Trump campaign – also segmented using psychometrics. The company amassed large amounts of data, built personality profiles for more than 100 million registered US voters and then, allegedly, used these profiles for targeted advertising.[52]

Another related trend is automatic content generation. Based on personal, psychological or other characteristics, personalised targeting can be used in combination with Natural Language Generation tools to automatically generate content for unique users. Dissemination of disinformation with aggressive automated methods just before the start of the campaign silence may adversely affect election results.

Although user profiling and political micro-targeting may simply be viewed as commercial advertising, these practices are problematic from a privacy and personal data protection point of view. The European Group on Ethics in Science mentions the right "to not be profiled, measures, analysed, … or nudged."[53] While users may believe that the encountered information is objective, spontaneous, citizen-generated, and universally encountered by other users, it is algorithms that decide what political views and information users come across online.[54] Relying on the collection and manipulation of users' data in order to anticipate and influence voters' political opinions and election results, user profiling and micro-targeting may pose a threat to democracy, public debate, and voters' choices.[55]

This takes us back to the underlying process of amassing and processing of vast amounts of personal data. Such data is often stripped of its original purpose(s) and may be used for objectives the individual is largely unaware of – in this case, profiling and targeting with political messages – in contravention of existing EU data protection principles.[56]

The EU has attempted to regulate the ways in which users' data is collected, stored, and used through its flagship data protection legislation, the General Data Protection Regulation (GDPR). In particular, Article 22 of the GDPR governs automated decision-making including detection models, assessment,

52 Michael Wade, "Psychographics: The Behavioural Analysis That Helped Cambridge Analytica Know Voters' Minds."

53 Judit Bayer et al., "Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and Its Member States" (Brussels: European Parliament, February 2019).

54 Ibid., 74.

55 Shara Monteleone, "Artificial Intelligence, Data Protection and Elections" (European Parliamentary Research Service (EPRS), May 2019).

56 Bayer et al., "Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and Its Member States," 75.

and automated profiling. This gives users the right not to be subject to a decision solely based on automated processing, including profiling (the so-called opt-out option). EU policymakers have argued that under the privacy law's requirements, personal data processed through automated decision-making cannot be used in political targeting.[57] In addition, Article 5 of the GDPR requires organisations to minimise the amount of data collected, and to restrict its use to its original intended purpose. Also noteworthy are Articles 13 and 15, according to which data subjects have a right to "meaningful information about the logic involved" and "the significance and the envisaged consequences" of automated decision-making.[58]

Two challenges are worth mentioning in this regard. First, Article 22 includes a number of requirements that could limit automated decision-making and profiling for companies using AI systems. For instance, the growing sophistication and complexity of algorithms make it challenging for companies to comply with the requirement of explainability. Algorithmic decision-making and behavior are difficult to explain and predict scientifically even for developers, yet according to the GDPR, companies must provide users with information that explains this in "clear and plain language" that is "concise, transparent, intelligible and easily accessible."[59] Second, AI systems need large training datasets to improve in accuracy and performance. Data minimisation, envisaged under Article 5, may limit access to training data, which would impact the AI system's ability to improve and more effectively tackle online threats, including disinformation.

3.2 *Deep Fakes*

Application of AI to audio and video content production presents an even bigger challenge. The so-called 'deep fakes' – digitally manipulated audio or visual material that is highly realistic and virtually indistinguishable from real material – were initially used in the movie industry. Nowadays, they are finding their application in the online realms of entertainment, consumer deception,

57 Gabriela Bodea et al., "Automated Decision-Making on the Basis of Personal Data That Has Been Transferred from the EU to Companies Certified under the EU-U.S. Privacy Shield" (European Commission, Directorate-General for Justice and Consumers, October 2018).

58 "Article 13: EU GDPR," PrivazyPlan, 2018, http://www.privacy-regulation.eu/en/article-13-information-to-be-provided-where-personal-data-are-collected-from-the-data-subject-GDPR.htm; "Article 15: EU GDPR," PrivazyPlan, 2018, http://www.privacy-regulation.eu/en/article-15-right-of-access-by-the-data-subject-GDPR.htm.

59 "Article 12: EU GDPR," PrivazyPlan, 2018, http://www.privacy-regulation.eu/en/index.htm.

and even politics and international affairs.[60] Commercial and even free software are already available in the open market. It is expected that soon, the only practical constraint on one's ability to produce a deep fake will be the availability of, and access to, a sufficiently large training dataset – i.e., video and audio of the person to be modeled.[61]

Forged videos and imagery still exhibit many artefacts which make them easy to recognise. By 2030, however, deep fakes could become indistinguishable from genuine information and easier to produce. Telling the difference between the original and manipulated content may become close to impossible for news consumers, and progressively difficult for machines.[62]

Highly realistic and difficult-to-detect depictions of real people doing or saying things they never said or did could discredit leaders and institutions, incite violence and tilt cities towards civil unrest, exacerbate existing divisions in society, or influence the outcome of elections.[63] The growing ease of making and sharing fake video and audio content across computers and mobile devices may create ample opportunities for intimidation, blackmail, and sabotage beyond the realm of politics and international affairs.[64]

There have been several instances of AI-generated videos depicting politicians making statements they never declared. In 2017, for example, computer scientists from the University of Washington produced a fake video of former US President Barack Obama to demonstrate a program they had developed, capable of turning audio clips into a realistic, lip-synced video of Obama speaking those words.[65] Although the Obama video was only a demonstration of how deep fake technology might be used, in May 2019, US House Speaker Nancy Pelosi was herself the victim of a deceptive video, in which she appears

60 Bayer et al., "Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and Its Member States."

61 Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics," *Foreign Affairs*, February 2019, https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war.

62 Bayer et al., "Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and Its Member States."

63 Robert Chesney and Danielle K. Citron, "Disinformation on Steroids: The Threat of Deep Fakes," Council on Foreign Relations (CFR), October 16, 2018, https://www.cfr.org/report/deep-fake-disinformation-steroids.

64 Chesney and Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics."

65 Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Schlizerman, "Synthesizing Obama: Learning Lip Sync from Audio," *ACM Transactions on Graphics* 36, no. 4 (July 2017).

to drunkenly slur her words. Although the video did not classify as a deep fake, it went viral on social media, prompting speculations about Pelosi's health condition.[66]

More recently, researchers at Global Pulse, an initiative of the United Nations (UN), devised a method to train AI to create fake UN speeches. They used a readily available language model (AWD-LSTM) trained on text from Wikipedia, and fine-tuned it on a dataset of UN General Assembly speeches. Within thirteen hours, the AI model was able to produce realistic speeches on a wide variety of debated topics, including climate change, immigration, and nuclear disarmament.[67] The experiment was intended to demonstrate the ease and speed with which the AI can generate realistic content, as well as the threat posed by a combination of this technique with other technologies, such as deep fakes.

> "*For instance, one may generate controversial text for a speech supposedly given by a political leader, create a 'deep fake' video of the leader standing in the UN General Assembly delivering the speech* (*trained on the large amount of footage from such speeches*), *and then reinforce the impersonation through the mass generation of news articles allegedly reporting on the speech*."[68]

Deep fakes make it possible for malign actors to deny the truth in two ways: not only may fake videos be passed off as real to create doubt but authentic information can be passed off as fake.[69] As the public becomes more educated about the threats posed by deep fakes, the latter technique is likely to become more plausible.

3.3 *Humans "out of the loop" of AI systems*

Although fully automated fact-checking remains a distant goal, as training datasets get bigger, AI systems will improve and can eventually replace human

66 Donie O'Sullivan, "Doctored Videos Shared to Make Pelosi Sound Drunk Viewed Millions of Times on Social Media," CNN, May 24, 2019, https://edition.cnn.com/2019/05/23/politics/doctored-video-pelosi/index.html.

67 Joseph Bullock and Miguel Luengo-Oroz, "Automated Speech Generation from UN General Assembly Statements: Mapping Risks in AI Generated Texts" (International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019).

68 Ibid.

69 Paul Chadwick, "The Liar's Dividend, and Other Challenges of Deep-Fake News," The Guardian, July 22, 2018, https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin.

oversight. Bots can amplify content but cannot create it yet. As the next wave of AI research focuses on creating tools that are better able to understand human language, context, and reasoning,[70] AI-enabled bots could end up in the driver's seat, with an ability to generate content, persuade, and tailor content for different audiences.[71]

There are legal reasons why humans need to be kept in the loop for content moderation. According to a recent study funded by the European Science-Media Hub, "limiting the automated execution of decisions on AI-discovered problems is essential in ensuring human agency and natural justice: the right to appeal. That does not prevent the suspension of bot accounts at scale, but ensures the correct auditing of the system processes deployed."[72]

The European data protection framework – which includes the GDPR – allows people to know how organisations are using their data as well as to contest certain decisions made by algorithms. Because developers cannot explain how algorithms produce certain outcomes (see previous section), complaints relating to the GDPR have already been lodged, several organisations have been sanctioned, and more cases are expected to follow.[73] From May 2018 until May 2019, European Data Protection Authorities (DPAs) received a total of 89,271 data breach notifications from companies, and 144,376 complaints from users.[74] Having humans in the loop, especially for judgment calls that impact other people's freedom, can help question the algorithm's decision as well as

70 Venkat Srinivasan, "Context, Language, and Reasoning in AI: Three Key Challenges," MIT Technology Review, October 14, 2016, https://www.technologyreview.com/s/602658/context-language-and-reasoning-in-ai-three-key-challenges/.

71 Rosenbach and Mansted, "Can Democracy Survive in the Information Age?"

72 Marsden and Meyer, "Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism," 16.

73 Adam Janofsky, "Large GDPR Fines Are Imminent, EU Privacy Regulators Say," The Wall Street Journal, May 3, 2019, https://www.wsj.com/articles/large-gdpr-fines-are-imminent-eu-privacy-regulators-say-11556829079; Stephanie Bodoni and Natalia Drozdiak, "U.S. Tech Giants Risk Hefty Fines, Irish Privacy Chief Warns," Bloomberg, June 12, 2019, https://www.bloomberg.com/news/articles/2019-06-12/european-regulator-probing-facebook-calls-for-global-data-laws; Davinia Brennan, "GDPR Enforcement Action – Polish & Danish DPAs Issue Their First Fines," *Lexology* (blog), April 26, 2019, https://www.lexology.com/library/detail.aspx?g=99bdf9be-2efe-49c9-a37a-75091c8f6b97.

74 "GDPR in Numbers" (European Commission, May 25, 2019), https://ec.europa.eu/commission/sites/beta-political/files/infographic-gdpr_in_numbers_0.pdf; Osborne Clarke, "GDPR One Year on: How Are EU Regulators Flexing Their Muscles and What Should You Be Thinking about Now?," Osborne Clarke, May 10, 2019, https://www.osborneclarke.com/insights/gdpr-one-year-eu-regulators-flexing-muscles-thinking-now/.

to proactively scrutinise the design, development, deployment, and use of AI applications – and to apply corrective measures when necessary.[75]

4 Solutions and Recommendations

Policymakers and politicians, user communities, fact-checkers, social media platforms, journalists, and all other stakeholders are grappling with a complex challenge which cannot be solved by a one-size-fits-all, single solution. For the legislator, the use of AI to counter disinformation and other online threats raises a host of regulatory questions. The following section outlines technical, legal, regulatory, and educational approaches – some existing, others emerging – that can help mitigate the challenges posed by AI systems in the context of information operations.

4.1 *De-emphasise and Correct False Content*

Social media companies can update their news feed algorithms to De-emphasise disinformation. In addition to flagging and downgrading false content, it is important for platforms to show effective corrections of verifiably false or misleading content that appeared online.[76] Of equal relevance is the dissemination of fact-based counter-messages. Although attribution in online disinformation campaigns is complicated, where sufficient evidence is available, it is important to publicly denounce the perpetrators of disinformation as well as to coordinate attribution and response.

4.2 *Promote Greater Accountability and Transparency*

Possible biases in algorithmic decision systems could be offset by the auditing of AI systems. Auditing would increase scrutiny of the data and the processes used to generate models using the data. A notable example in this regard is the *Algorithmic Accountability Act*, a draft regulation recently proposed by the US that would require companies to audit their AI systems for bias and discrimination, issue impact assessment, and implement corrective

75 Olivier Panel, "Algorithms, the Illusion of Neutrality: The Road to Trusted AI," Medium, April 10, 2019, https://towardsdatascience.com/algorithms-the-illusion-of-neutrality-8438f9ca8471.

76 See, for example, a call issued by Avaaz for Facebook, Twitter and all technology platforms to issue corrections to fake news: "Facebook: Issue Corrections to Fake News!," Avaaz, February 12, 2019, https://secure.avaaz.org/campaign/en/correct_the_record_imp/.

measures.[77] Making training in ethics part of the computer science curriculum – i.e., teaching how to build "ethical by design" applications – could limit the probability of biases being fed into the codes.[78]

In addition to greater accountability, there are mounting calls for increased algorithmic transparency.[79] Such proposals have met with strong resistance from tech companies and developers, who argue that revealing the source code – i.e., the system's inner workings – would force them to disclose proprietary information and harm their competitive advantage.[80]

4.3 *Technological Remedies for Deep Fakes*

As regards solutions for countering deep fakes, law professors Robert Chesney and Danielle Citron propose three technological remedies. The first relates to enhanced detection of forged material using forensic tools. As most training datasets lack faces with eyes closed, techniques that look for abnormal patterns of eyelid movement have been developed to improve the detection of deep fakes. However, as deep fake technology evolves based on a virus / antivirus dynamic, once this forensic technique was made public, the latest generation of deep fakes adapted shortly after.[81]

The second technological solution involves the authentication of content before it spreads – the so-called digital provenance solution.[82] If audio, photo, and video content can be digitally watermarked at the moment of its creation, such credentials could later be used as a reference to compare to suspected fakes.[83] A third, more theoretical technological approach, revolves around "authenticated alibi services" that would monitor and store all individual's actions, movements, and locations in order to prove where one was and what

77 Adi Robertson, "A New Bill Would Force Companies to Check Their Algorithms for Bias," The Verge, April 10, 2019, https://www.theverge.com/2019/4/10/18304960/congress-algorithmic-accountability-act-wyden-clarke-booker-bill-introduced-house-senate.

78 Gitika, Women and the Fourth Industrial Revolution.

79 Bentzen, "Computational Propaganda Techniques."

80 See for example: Kartik Hosanagar and Vivian Jair, "We Need Transparency in Algorithms, But Too Much Can Backfire," Harvard Business Review, July 25, 2018, https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire.

81 James Vincent, "Deepfake Detection Algorithms Will Never Be Enough," The Verge, June 27, 2019, https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work.

82 Chesney and Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics."

83 Ibid.

he or she was saying or doing at any given time.[84] Although alibi services and enhanced lifelogging may be particularly valuable for high-profile individuals with fragile reputations, such as celebrities and politicians, they have serious negative implications for personal privacy.

Another proposal is to use the same tools that generate deep fakes to detect them. Karen Hao of the MIT Technology Review recommended that governments require companies and researchers who produce tools for deep fakes to invest in countermeasures, and that social media and search companies integrate those countermeasures directly into their platforms.[85]

4.4 *Regulate Social Media Content?*

European and American policymakers are grappling with the possibilities to regulate online content. Existing proposals have either placed additional responsibility and liability on platforms or provided governments with more control over online content. The proposed rules have raised a number of challenges and have been met with resistance from various stakeholders – platforms, civil rights organisations, and end users alike.

Social media platforms have deployed technical tools and other capabilities to address disinformation through self-regulation and R&D investment. Following the European elections of May 2019, a number of organisations and policymakers have argued that self-regulation efforts do not suffice.[86] In their views, as online platforms have significant civic power and control over data, the major role they play in privacy protection and content moderation should remain subject to enforceable regulation, external oversight, and independent impact assessment to ensure compliance with fundamental rights.

In contrast, others view audited co-regulation as a more desirable governance system in that it is more fit for this era and context, and for the sheer size and speedy evolution of the problem. Proponents of such protocols consider that the EU Code of Practice on Disinformation has set an example on how governments and civil society can work with industry in the digital economy, act in coordination with technology experts to tackle complex issues, and address the challenges of evolving technologies while harnessing their benefits.[87]

84 Chesney and Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics."

85 Karen Hao, "Deepfakes – What Needs to Be Done Next?" (Academia, June 12, 2019), https://www.academia.edu/39586141/Deepfakes-_What_needs_to_be_done_next.

86 Marietje Schaake, "Letter Calling for Parliamentary Inquiry Tech Companies – Democracy," Marietje Schaake, June 5, 2019, https://marietjeschaake.eu/en/letter-calling-for-parliamentary-inquiry-tech-companies-democracy.

87 "Code of Practice against Disinformation: Commission Recognises Platforms' Efforts Ahead of the European Elections" (European Commission, May 17, 2019), https://europa

One critical aspect of the solution lies in the determination of roles, responsibilities, and liability of the various stakeholders involved. Rebalancing this ecosystem should not mean designating online platforms as both judges and jury in determining what truth is. This could lead to overcensorhip, as out of an abundance of caution and by fear of penalties, platforms could remove lawful content.[88] This risk raised the controversy over the German law on fake news, in force since 1 January 2018. The law imposes a 24-hour timeframe under which platforms have to take down fake news and hate speech – a constraint which is impractical for large platforms, and thus unworkable for smaller companies.[89] In addition, as smaller platforms have limited resources, it is not realistic to expect them to police the entire content. This holds true for bigger Internet companies as well: while Facebook's user community is larger than the populations of China and India, the number of its employees working on safety and security of online content barely matches that of Belgium's police forces – 30,000 people.[90] A more efficient framework to regulate the online environment in the context of polluted content requires more support for those social media firms which are grappling with an issue that has become bigger than themselves.

Governments should not be those solely in charge of monitoring online content either. They often lag behind the private sector in terms of technical expertise, infrastructure, and their understanding and evaluation of new technologies.[91] Often, technologies evolve far quicker than government policies, which can rapidly become obsolete. In addition, governments are not neutral data brokers either. Increasing governmental power over data tends to raise concerns over statutes that could be used to infringe on civil liberties and

.eu/rapid/press-release_STATEMENT-19-2570_en.htm; Chris Marsden and Trisha Meyer, "Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism" (European Parliamentary Research Service (EPRS), March 2019).

88 Darrell M. West, "How to Combat Fake News and Disinformation" (The Brookings Institution, December 18, 2017), https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/.

89 Patrick Evans, "Will Germany's New Law Kill Free Speech Online?," *BBC Trending* (blog), September 18, 2017, https://www.bbc.com/news/blogs-trending-41042266.

90 Katie Harbath and Samidh Chakrabarti, "Expanding Our Efforts to Protect Elections in 2019," Facebook Newsroom, January 28, 2019, https://newsroom.fb.com/news/2019/01/elections-2019/.

91 "Algorithmic Accountability Policy Toolkit" (AI Now Institute, October 2018), https://ainowinstitute.org/aap-toolkit.pdf; Tam Harbert, "Can the Government Keep Up with the Pace of Tech?," Techonomy, November 11, 2018, https://techonomy.com/2018/11/can-government-keep-pace-tech/.

give governments the ability to suppress the freedom of speech for political purposes.

For instance, Singapore's recent draft law presented a clear threat to free expression and freedom of the press, by allowing law ministers to decide without judicial review whether online content (described as "factual information") is true or false.[92] In addition, these rules would allow the government – rather than judges – to forbid statements aiming to "diminish public confidence" in Singaporean state institutions. Such language creates legal uncertainty and leaves much room for interpretation, potentially stifling freedom of speech.[93] Similarly, a recent Russian legislation criminalised the spread of online disinformation, including statements that "disrespect" the state. Such remarkably vague language could enable political censorship to silence opponents.[94]

It is important to note that attempts to regulate and devise policy for a technology whose definitions, risks, challenges, and contexts vary require caution and a constant dialogue with all stakeholders. This should be a cooperative venture from industry, academia, and government, and the typical regulatory approach will not necessarily work in a fast-moving environment. The diversity of online companies calls for a variety of adequate rules and standards for accountability. Suggesting a uniform implementation of one-size-fits-all requirements would be misguided.

Efforts should also be more inclusive. For instance, the Code of Practice on Disinformation only included a handful of major online platforms while falsified content travels and migrates to many others, such as 4Chan, 8Chan or Reddit (where "Pizzagate"[95] started). What is more, the Code only focused on transparency for political advertisements, and only in Europe. Overall, not all stakeholders are convinced that this policy framework has produced impactful, satisfactory results.

92 "Protection from Online Falsehoods and Manipulation Bill," Pub. L. No. 10/2019 (2019), https://www.parliament.gov.sg/docs/default-source/default-document-library/protection-from-online-falsehoods-and-manipulation-bill10-2019.pd.

93 Karishma Vaswani, "Concern over Singapore's Anti-Fake News Law," BBC News, April 4, 2019, https://www.bbc.com/news/business-47782470.

94 "Putin Signs 'Fake News,' 'Internet Insults' Bills Into Law," The Moscow Times, March 18, 2019, https://www.themoscowtimes.com/2019/03/18/putin-signs-fake-news-internet-insults-bills-into-law-a64850.

95 "Pizzagate" is the name of a conspiracy theory popularised on social media platforms such as 4Chan and Reddit during the 2016 US presidential election campaign by opponents of Hillary Clinton's candidacy. As it eventually escalated to criminal reactions including a shooting, "Pizzagate" is often referred to as an example of how disinformation can have dire consequences.

4.5 *TechPlomacy Alongside Diplomacy*

Following Denmark's lead, countries can increase the engagement of and trust among different stakeholders by setting up "tech delegations" or "tech ambassadors", or by assigning some of these responsibilities to existing, relevant national authorities. Recognising that technology companies are important actors influencing global policies, techplomacy holds the potential for creating new avenues for dialogue and collaboration between technology industry and governments.[96] Countries' decision-makers can use this relationship to discuss issues such as election meddling, disinformation and harmful content, cybersecurity, or the collection of e-evidence for policy investigation.[97] Techplomacy can also help ensure that tech companies step up to the plate and assume a responsibility that is proportional to the kind of influence they wield.[98] At the national level, techplomacy can be assigned to a person or a body responsible for overseeing and coordinating efforts of existing cyber envoys and diplomats. To ensure it can effectively uphold such responsibilities, this national authority should be given a clear mandate and a strong political backing.

4.6 *Break up Big Tech?*

Work and action from online platforms, fact-checkers, and governments are necessary but, if isolated, these actors' efforts will not be powerful enough to contain disinformation. US Senator Elizabeth Warren and German Member of the EU Parliament Katarina Barley recently called for "breaking up" large technology companies – a move that would make it more costly for malicious actors to use multiple channels to propagate disinformation.[99]

A push to dismantle big tech may have counterproductive consequences. Social media platforms' current solutions may not be perfect but it is their large base of users that makes them the most accurate and impactful. In addition, multiple platforms – rather than a few – could make it more difficult to

96 "TechPlomacy," Office of Denmark's Tech Ambassador, n.d., http://techamb.um.dk/en/techplomacy/.

97 Tom Foremski, "The First Ambassador to Silicon Valley Struggles with 'TechPlomacy,'" ZDNet, February 1, 2019, https://www.zdnet.com/article/danish-ambassador-to-silicon-valley-struggles-with-techplomacy/.

98 Sean Brocklehurst, "Dane against the Machine: Tech-Diplomat Aims to Protect Fundamentals of Democracy in Digital Age," CBC News, February 24, 2019, https://www.cbc.ca/news/technology/national-casper-klyge-tech-ambassador-1.4828015.

99 Georg Ismar, Mathias Mullen von Blumencron, and Sonja Alvarez, "'We Are Talking about Breaking Monopolies like Facebook,' Says Barley, Who Tops SPD's EU Election List," Euractiv, April 18, 2019, https://www.euractiv.com/section/copyright/interview/we-are-talking-about-breaking-monopolies-like-facebook-says-barley-who-tops-spds-eu-election-list/.

adequately address disinformation, as responses would be more fragmented, while propellers of disinformation can easily replicate their actions on several platforms.[100]

4.7 *Media and Digital Literacy*

In the fight against disinformation, technological solutions are not enough to combat the problem. As put by Dr. Alexander Klimburg, Director of the Cyber Policy and Resilience Program at The Hague Centre for Strategic Studies, "attacking the body of cyber (the technical layers) is just a detour to attacking the mind (the human being)."[101] Responses should therefore go beyond the *technical* and focus on the *psychological* dimension. Ultimately, the reason why disinformation works is because there is an audience for it.

Increasing media and digital literacy may be one of the most efficient and powerful tools to restore a healthy relationship to information and increase the resilience of our democracies to online disinformation. Digital and media literacy education should be encouraged from early childhood. The focus should not only be on children but also on election officials, elderly citizens, and marginalised and minority groups.[102] In fact, elderly citizens, who face the biggest gap in terms of digital literacy, are most likely to vote in national elections.[103]

Finland provides a good example to follow. Already in 2014, the government launched an anti-fake news initiative targeting residents, students, journalists, and politicians with the objective to teach how to counter false information designed to sow division.[104] The 2016 reform of the country's education system aimed to emphasise critical thinking. While a number of European countries have launched anti-disinformation campaigns at schools, Finland's program also teaches more specialised skills and techniques, such as how to identify a troll or bot by taking a closer look at their social media profile.[105] Ahead

100 Michael R. Strain, "Breaking Up Facebook Would Make Things Worse," Bloomberg, July 1, 2019, https://www.bloomberg.com/amp/opinion/articles/2019-07-01/facebook-breakup-would-worsen-privacy-problem?__twitter_impression=true.

101 Alexander Klimburg, *The Darkening Web: The War for Cyberspace* (Penguin Press, 2017), 55.

102 Clara Tsao, "Disinformation, Global and Continental Case Studies" (May 29, 2019), https://www.disinfo.eu/2019/06/07/eu-disinfolab-annual-conference/.

103 Ibid.

104 Eliza Mackintosh, "Finland Is Winning the War on Fake News. What It's Learned May Be Crucial to Western Democracy," CNN, May 18, 2019, https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/.

105 Kristine Berzina et al., "The ASD European Policy Blueprint For Countering Authoritarian Interference in Democracies" (Washington DC: The German Marshall Fund of the United States (GMF), 2019).

of national elections held in April 2019, government-commissioned adverts warned against disinformation, encouraging voters to think independently.[106] Alongside the media campaign, the Finnish government provided anti-disinformation training to political parties and candidates alike, and Finnish fact-checking agency Faktabaari (FactBar) developed a digital literacy "toolkit" for students from elementary to high school level learning about EU elections.[107] As regards AI in particular, Finnish technology firm Reaktor and the University of Helsinki joined forces to teach various aspects of AI for free to anyone interested in the technology.[108]

Social media platforms are also investing in digital literacy initiatives. In 2018, Facebook launched a Digital Literacy Library in six languages to help young people consume information critically and produce and share content responsibly.[109] The same year, Twitter partnered with UNESCO to promote more media and information literate citizenry in online spaces.[110]

EU institutions have a key role to play in streamlining similar efforts, for instance by initiating information campaigns across Member States that raise awareness of social media and smart use of emerging technologies. The EU Media Literacy Week, which aims to underline the societal importance of media literacy and to promote media literacy initiatives and projects across the EU, is in this respect a step in the right direction.[111]

4.8 *Cybersecurity*

Malicious actors are increasingly merging disinformation with traditional cyber attacks. With growing frequency, social media platforms are the target of

106 "Election Ads Urge Finns 'Think for Yourself' amid Fake News Fears," France24, April 13, 2019, https://www.france24.com/en/20190413-election-ads-urge-finns-think-yourself-amid-fake-news-fears.

107 Mackintosh, "Finland Is Winning the War on Fake News. What It's Learned May Be Crucial to Western Democracy."

108 "Elements of AI: Finland Is Challenging the Entire World to Understand AI by Offering a Completely Free Online Course – Initiative Got 1% of the Finnish Population to Study the Basics," University of Helsinki, September 6, 2018, https://www.helsinki.fi/en/news/data-science-news/finland-is-challenging-the-entire-world-to-understand-ai-by-offering-a-completely-free-online-course-initiative-got-1-of-the-finnish-population-to.

109 Facebook, "Digital Literacy Library," Digital Literacy Library, n.d., https://www.facebook.com/safety/educators.

110 "UNESCO Partners with Twitter on Global Media and Information Literacy Week 2018," UNESCO, October 25, 2018, https://en.unesco.org/news/unesco-partners-twitter-global-media-and-information-literacy-week-2018.

111 "European Media Literacy Week," European Commission, March 25, 2019, https://ec.europa.eu/digital-single-market/en/news/european-media-literacy-week.

data breaches, malware attacks, network penetration, and social engineering. Advances in machine learning will enable adversaries to automate malware and offensive cyber capabilities, avoid detection, and evade defensive measures in place.[112]

Securing the digital infrastructures upon which governments, businesses, and wider society increasingly depend and educating citizens about personal cybersecurity is important to effectively fend off disinformation and related cyber threats. In addition to essential infrastructure, attention should also be paid to strengthening cybersecurity in electoral systems and processes. A combination of national legislation, industry action, and internationally agreed approaches can ensure that the necessary safeguards are in place.

4.9 *R&D for AI*

Getting ahead of disinformation attacks will require more investments in R&D for AI, in order to improve algorithms and their ability to detect false content. The EU can allocate more funding towards the intersection of AI and disinformation – beyond its current efforts. In 2019, the European Commission increased the budget for its External Action Service's strategic communications team from €1.9 million to €5 million, to support its mission to address disinformation and raise awareness about its adverse impacts.[113] Under the Horizon 2020 Programme, the Commission earmarked an additional €25 million for research and innovation projects that develop tools to identify content and analyse networks, and to better understand information cascades across various platforms. The EU also invested €1.5 million in the creation of prototypes such as pilot platforms to support and scale up cooperation between universities, researchers, and the fact-checking community. An additional €2.5 million was set apart for the creation of the Social Observatory for Disinformation and Social Media Analysis (SOMA), a secure platform to enhance collaboration across disciplines and specialisations.[114] In comparison, according to the 2015 estimates of the US Department of State, Russia invests as much as

112 Faesen et al., "Understanding the Strategic and Technical Significance of Technology for Security: Implications of AI and Machine Learning for Cybersecurity."

113 "Questions and Answers – The EU Steps up Action against Disinformation" (European Commission, December 5, 2018), http://europa.eu/rapid/press-release_MEMO-18-6648_en.htm.

114 Based on the intervention of Paolo Cesarini, "Using AI to Fight Disinformation in European Elections" (February 20, 2019), https://www.datainnovation.org/2019/02/event-recap-using-ai-to-fight-disinformation-in-european-elections/.

$1.4 billion a year on internal and external propaganda, reaching 600 million people across 130 countries.[115]

5 Conclusions

As the volume of online content continues to grow, automated fact-checking holds great potential as a speedier and cost-efficient complement to, or even replacement of, human oversight – by blocking or removing false content before it is uploaded online. It might take another five to ten years for AI to make nuanced distinctions and proactively identify harmful content embedded in linguistic, cultural, and political contexts with minimal to no human input.[116] For as long as AI does not grasp context and grey areas, human supervision remains critical. As regards their accuracy, detection algorithms need to be further developed to reach the efficiency level of e-mail spam filters.

As this paper demonstrates, the development of AI systems is a two-edged sword for democratic societies. On the one hand, AI systems will improve human processes and tasks in the online environment, such as detection of disinformation, bots, altered text and images, and manipulated audio and video material. On the other hand, when the same technologies are adopted by adversaries, they will enable them to magnify the effectiveness and scale of information operations. As ideological and geopolitical tensions between democratic and authoritarian states continue to grow, AI and computational propaganda are likely to become tools of political warfare used against democratic societies.

There needs to be a greater global effort to work on ways to detect and respond to AI-generated content. Policies aiming to combat false and harmful content should already be focusing on the next generation of disinformation which, fuelled by advances in AI and decentralised computing, promises to spread faster, to be more sophisticated, and harder to detect.

New technologies evolve far quicker than government policies and often undermine existing legal and policy frameworks. In order to ensure responsible use of AI, as well as to develop the right responses for its potential misuses early on, stronger connections, partnerships, and open conversations need to

115 Molly McKitterick, "Russian Propaganda: 'The Weaponization of Information,'" Voice of America (VOA), November 3, 2015, https://www.voanews.com/europe/russian-propaganda-weaponization-information.

116 Faesen et al., "Understanding the Strategic and Technical Significance of Technology for Security: Implications of AI and Machine Learning for Cybersecurity."

be established between policymakers, engineers, and researchers.[117] Acknowledging that technology companies, including social media platforms, can provide powerful solutions, governments and other stakeholders should strive to cooperate with them in order to develop better filters to prevent the spread of disinformation. Broader ex-ante consultation with online platforms, users, and other stakeholders would help prevent pitfalls such as unrealistic legislative proposals, a lack of balance in the distribution of responsibilities, and regulation or infringement of freedom of expression. Enhancing dialogue between relevant stakeholders will generate more realistic and agile policies.

In parallel, there needs to be more research aimed at understanding the scale, scope, and origin of disinformation, the trends and patterns behind it, and the mechanisms used by malicious actors (both state and non-state) to organise their actions and amplify disinformation. Investigating the veracity of content, the information cascade, and the spread of disinformation requires more time, more research – hence more funding – better tools, more neutral algorithms, but also greater access to data for independent researchers. Regarding the latter, publishing datasets comes with a number of challenges and concerns, including those holding to privacy, the difficult conversion of datasets into actionable information, and potential misuse of data by malign actors.

This paper explored a number of worthy approaches – some already existing and others yet to emerge – that can help mitigate the challenges posed by AI systems in the context of disinformation campaigns. Many of the proposed solutions bring challenges of their own. In the fight against disinformation, there is no single fix. The next wave of disinformation calls first and foremost for societal resilience. Investing in digital and media literacy in a bid to enhance societal awareness and to increase critical media consumption is essential. To preserve democratic values and the stability of societies, governments, media, and the private sector need to work together to share best practices and develop tools that will provide durable and sustainable solutions in the future.

There is a broader role for international organisations in building policies for AI as well as societal resilience against disinformation, including by monitoring and informing about the uses and applications of AI systems. In addition to building awareness among the general public about the problem, and promoting digital and media literacy across the OSCE area, the OSCE could encourage greater information sharing about disinformation campaigns and collaboration among all relevant stakeholders in the OSCE area. The OSCE

117 Steven Feldstein, "We Need to Get Smart About How Governments Use AI," Carnegie Endowment for International Peace, January 22, 2019, https://carnegieendowment.org/2019/01/22/we-need-to-get-smart-about-how-governments-use-ai-pub-78179.

could also join forces with other international organisations (namely the EU) to develop guidelines for the ethical development and use of AI systems across the OSCE area. Additional funding and support should be directed towards independent and automated fact-checking initiatives, academic research on AI-powered disinformation, innovation, and cross-border and cross-sector knowledge transfer.[118]

118 For a more detailed assessment of the ways in which the OSCE can leverage the opportunities presented by new technologies see the OSCE Perspectives 20–30 report (forthcoming).