Policy Brief

Ensuring Ethical AI Practices to counter Disinformation

Luiss DataLab
Federica Urzo, Elisabetta Panico, Salvatore Custureri
**Disclaimer**
This document was written as part of the MediaFutures project under EC grant agreement 951962. The information, documentation and figures available in this deliverable were written by the MediaFutures project consortium and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

**Statement of originality**
This document contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

**How to quote this document**

This document is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).
INTRODUCTION

Disinformation has become a critical issue in today’s interconnected world, with deliberate spread of false or misleading information posing significant threats to democracy, social cohesion, and public trust. The rise of digital platforms and social media has worsened the problem, making it harder to distinguish truth from falsehood and amplifying the impact of disinformation. This policy brief aims to provide a comprehensive understanding of disinformation and offer policy recommendations to mitigate its harmful effects. The consequences of disinformation are wide-ranging, undermining public discourse, trust in media, and fostering polarisation. Disinformation can manipulate public opinion, influence elections, and incite violence. The COVID-19 pandemic and Russia-Ukraine War have highlighted the dangers of health-related disinformation, jeopardising public health efforts. Addressing disinformation requires a collaborative approach involving governments, technology companies, civil society, and individuals. Policy interventions at all levels should focus on promoting media literacy, transparency, accountability, international cooperation, and supporting fact-checking organisations. Balancing countering disinformation, while protecting freedom of expression, is crucial. Measures must avoid suppressing legitimate speech and adhere to principles of transparency, accountability, and human rights norms and provisions. In the context of the Russia-Ukraine conflict, disinformation has been extensively used, facilitated by social media and the internet. False narratives, conspiracy theories, and historical revisionism have distorted facts and manipulated public opinion. Online platforms and algorithms contribute to the spread of disinformation by creating echo chambers and confirmation biases. Efforts have been made at the EU and International level to combat disinformation in providing accurate information, and supporting independent media, while sanctioning state-linked media, and improving fact-checking programs.

Yet, striking a balance between countering disinformation and preserving freedom of expression including press freedom remains a challenge. Effective strategies are needed to counter disinformation while strengthening information ecosystems and democracy. In recent years, there has been a proliferation of disinformation, fueled by the consumption of news and information from unverified and unreliable sources online. Moreover, the emergence of deep fakes has raised concerns about their potential misuse. These are highly realistic digitally manipulated media created thanks to artificial intelligence (AI) software that, starting from real content (images and audio), can modify or recreate, in an extremely realistic way, the features and movements of a face or body and faithfully imitate a given voice.
The emergence of deepfake technology has garnered significant attention from researchers, policymakers, and the public.

These manipulated media, including videos, audio, photos, and text, raise concerns about potential misuse, such as extortion, harassment, and misinformation. The impact of deepfakes extends beyond direct abuse, as they have the potential to erode trust in digital content, leading to a destabilising effect on society. The increasing quality, accessibility, and affordability of deepfake technology contribute to its widespread use and distribution. Additionally, societal shifts towards user-generated content and the dominance of imagery in online communication further amplify the significance of deep fakes. In the context of AI it is becoming a threat also the voice cloning technology replicates human voices using Concatenative TTS and Parametric TTS approaches. Voice cloning is connected to text synthesis, generating content that matches the target’s style and vocabulary.

Moreover, The Europol report provides a detailed overview of the criminal use of deepfake technology. Participants in the foresight activities highlighted trends that European law enforcement agencies should pay attention to. Notably, “crime as a service” (CaaS) was mentioned, where criminals sell access to tools, technologies, and knowledge for cyber and technology-enabled crimes. CaaS is expected to evolve alongside current technologies, automating crimes like hacking and adversarial uses of machine learning and deepfakes. Criminals tend to adopt new technologies early, putting them ahead of law enforcement in implementation and adaptation. The increasing availability of disinformation and deepfakes will profoundly affect how people perceive authority and information. The rise of deepfakes undermines trust in authorities and official facts, potentially leading to a reality where citizens lack shared truths, termed “information apocalpyse” or “reality apathy.”

Recognizing and preparing for this manipulation is crucial to differentiate between benign and malicious uses of this technology. The “Malicious Uses and Abuses of Artificial Intelligence” report by Europol, TrendMicro, and UNICRI includes a case study on this. The report highlights how deepfake technology can facilitate various criminal activities such as harassment and humiliation online, extortion and fraud, facilitating document fraud, faking online identities and fooling verification mechanisms, non-consensual pornography, online child exploitation, falsifying electronic evidence for criminal investigations, disrupting financial markets, distributing disinformation and manipulating public opinion, supporting extremist or terrorist group narratives, stirring social unrest and political polarization. Moreover Businesses are also vulnerable to disinformation campaigns, as deepfakes can be employed to create false information that deceives the public. For instance, malicious actors could produce deepfakes that falsely depict a company executive engaging in controversial or illegal activities. Some deepfakes could be used for deceptive advertising and spreading false information, leading to negative publicity for the targeted company. These applications of deepfakes could affect domains like the stock market and company valuation, as the public (including stakeholders, shareholders, and consumers) might believe the fabricated content, prompting them to sell stocks or boycott the company.

**WHAT ARE THE RISKS AND CHALLENGES CONNECTED TO AI**

AI-powered deepfake technology has become more proficient at creating realistic videos and audios of specific individuals. This poses a significant risk as malicious actors can use deepfakes to spread fake news, launch cyber attacks, and harm reputations:

**Political Manipulation**: Deep fakes have the potential to disrupt electoral processes by spreading fabricated videos or audio to manipulate public opinion. Policymakers must be prepared to address this threat to democratic processes.

**Trust and Authenticity**: Deep fakes erode trust in media and public figures by blurring the lines between reality and fiction. Ensuring the authenticity of information is crucial in maintaining public confidence.
Legal and Ethical Implications: Intellectual property, fair use exemptions, data protection and consent issues arise in the creation and dissemination of deep fakes. Legal frameworks need to adapt to address these challenges while safeguarding individual rights.

However, technologies and in particular the use of AI can offer significant advantages in various domains, including education. Policymakers should consider these potential benefits while formulating policies:

Education: Compelling Teaching Tools. Artificial Intelligence like deep fake can enhance educational experiences by creating engaging videos featuring historical figures or modifying existing content to illustrate pedagogical points.

Affordable Production: AI technology enables cost-effective creation of video content and images that can be tailored to specific educational or entertainment scenarios, expanding access to immersive learning experiences.

Art: Creative Expression: Generative AI and Deep fakes offer artists a medium for expressing ideas, satirising public figures, and critiquing societal issues. They can push the boundaries of artistic expression and foster critical thinking.

Activism and Social Commentary: Deep fakes can be used by activists to convey their messages in impactful ways, complementing traditional forms of advocacy.

Evolving Techniques: As deepfake detection tools improve, so do the techniques used by malicious actors. The same AI technologies that are used to detect deepfakes can also be employed to train the next generation of more sophisticated deepfakes, creating a cat-and-mouse game between detection and creation.

Collaboration and Standards: Organizations and industry bodies are working together to develop technical standards and best practices for verifying the authenticity of content. Initiatives like the Coalition for Content Provenance and Authenticity (C2PA) aim to establish guidelines and promote collaboration among companies to tackle the deepfake challenge.

Security Measures: Strengthening overall cybersecurity measures can help mitigate the risks associated with deepfakes. Implementing multi-step authentication processes, adopting industry related norms and new standards, and staying updated on the latest security technologies can enhance defenses against deepfake attacks.

Deepfake Detection: Tools and techniques for detecting deepfakes are also advancing. Deepfake detectors can search for biometric signs within a video or audio that indicate manipulation. Improvements in detection methods can help identify and mitigate the spread of deepfakes.

While the use of AI in the fight against disinformation has risks, such as the increasing sophistication of deepfake technology, there are also opportunities to develop better detection tools, raise awareness, foster collaboration, and enhance security measures to counter the threats posed by deepfakes.

RECOMMENDATIONS AND PROPOSED SOLUTIONS FOR PLATFORMS AND CIVIL SOCIETY

This chapter proposes solutions to address disinformation and new technologies. The collaboration between platforms, governments, and civil society, promotes transparency in platform policies, effective content moderation, the promotion of fact-based information, and compliance with laws. Investment in advanced technologies to detect deep fakes and advocates for public awareness through media literacy education and cooperation with tech companies to combat the spread of deep fakes. In particular we recommends to
**Strengthen Collaboration:** Platforms should adopt a collaborative approach involving various stakeholders, including governments, civil society organisations, and fact-checkers, to counter the spread and impact of disinformation. This can include sharing information, best practices, and resources to develop effective strategies.

**Enhance Transparency:** Platforms should prioritise transparency by providing clear and comprehensive information on their policies, algorithms, and content moderation processes. Users should have a better understanding of how their data is used, and how algorithms work to prevent the amplification of false and misleading narratives.

**Implement Effective Content Moderation:** Platforms need to allocate sufficient resources to effectively monitor and moderate harmful content. This includes investing in advanced AI systems and human moderation teams to detect and remove disinformation in a timely manner. Transparent and consistent guidelines should be in place to ensure fairness and accountability in content moderation decisions.

**Promote Fact-based Information:** Platforms should prioritise the promotion of fact-based information from reliable sources. This can be done by partnering with credible news organizations and fact-checkers to provide accurate information and combat false narratives.

Advertising promoting climate change denial or other forms of misinformation should be prevented.

**Improve the Access to Data for Researchers:** Platforms should make efforts to provide access to data for independent researchers to evaluate the effectiveness of their policies and initiatives in countering disinformation. This will enable better analysis and understanding of the impact of disinformation and the effectiveness of countermeasures.

**Comply with Regulatory Frameworks:** Platforms should fully comply with regulatory frameworks, such as the Digital Services Act (DSA) or other relevant International, EU and National laws and regulations, that provide for obligations on addressing disinformation and mitigating associated risks, the Code of Practice on Disinformation that aims to commit signatories to a range of actions to counter disinformation. These actions include providing transparency reports on political advertising, restricting advertising placements on disinformation websites, disrupting advertising revenue for purveyors of disinformation, and enabling user feedback and fact-checking mechanisms. In this framework, compliance should not be limited to large platforms but extended, with adjustments, to smaller platforms to ensure a comprehensive approach.

**Strengthen Oversight:** Platforms should proactively cooperate with regulatory authorities to enhance oversight. This includes addressing specific issues related to platforms like Twitter, Reddit, Telegram, Spotify, and TikTok mentioned in the report. Platforms should take swift action to address shortcomings, such as reinstating accounts involved in spreading disinformation or allowing abusive content.

**Invest in Responsible AI Development:** Platforms should prioritise the responsible development and deployment of AI technologies. This involves ensuring that AI systems, including large language models like ChatGPT, are designed and used with respect for human rights, ethical considerations, and the rule of law. Regular audits and evaluations of AI systems should be conducted to minimise the risks associated with the spread of propaganda and disinformation.

**Foster International Cooperation:** Platforms should actively engage in international cooperation efforts, such as the Coordinated EU strategy against foreign information manipulation interference. Collaboration with different institutions and organisations across borders will help monitor and counteract disinformation campaigns that seek to undermine democratic processes.
Technological Solutions: Investing in research and development of advanced algorithms to detect deep fakes can aid in identifying and mitigating their harmful effects. Implementing standards for content attribution and labelling can enable users to distinguish between real and synthetic media. Strengthening copyright laws to protect against unauthorised use of individuals’ likenesses or intellectual property in deep fakes.

Media Literacy and Awareness: Fact-Checking and Verification: Promoting media literacy initiatives to educate the public about deep-fake technology, encouraging critical thinking, and providing tools for fact-checking. Collaboration with Tech Companies: Encouraging responsible practices among tech companies and social media platforms to combat the spread of deep fakes.

THE ROLE OF TECHNOLOGY COMPANIES IN COUNTERING DISINFORMATION

The role of technology companies in combatting disinformation is paramount, as they bear the responsibility of upholding content integrity and reliability on their platforms. In recent years, driven in part by the European Union’s scrutiny, there has been mounting pressure on tech firms to take proactive measures against the scourge of disinformation.

To fulfill this duty, tech companies employ teams of moderators tasked with reviewing user-generated content, promptly removing or labeling material that contravenes their community guidelines or terms of service. These companies leverage automated systems and artificial intelligence algorithms to identify and flag potentially false or misleading information. Moreover, they establish partnerships with fact-checking organizations to verify content accuracy, equipping these fact-checkers with tools and data to assess information veracity and flag any inaccuracies.

Another crucial facet is the continuous refinement of algorithms. Tech companies consistently need to update their algorithms to prioritize trustworthy and reliable information sources. This effort aims to reduce the visibility of false or substandard content in users’ feeds and search results, thereby enhancing the quality of the information ecosystem.

To engage users in the fight against disinformation, tech companies need to provide reporting tools, allowing users to flag potentially misleading or false content. These reports undergo thorough review, leading to appropriate actions.

Furthermore, some tech firms publish regular transparency reports detailing their content moderation efforts, including the removal of false or misleading information. These reports serve to enhance transparency and accountability.

Recognizing the importance of education, tech companies invest in programs and initiatives that promote media literacy and critical thinking skills among users. The goal is to empower individuals to discern and evaluate reliable information sources.

Finally, collaboration is key. Tech companies must actively engage with researchers, academics, and experts in the field of disinformation to gain a deeper understanding of the problem and to jointly develop effective strategies to combat it.
The role of government in countering disinformation, including the development of legal frameworks and the allocation of resources to combat disinformation.

Ensuring a strong understanding of online disinformation also involves authorities staying informed about the latest trends and narratives. Monitoring disinformation can be highly resource-intensive due to the vast amount of online information and social media content being shared. Given that many Local Regulatory Authorities should receive adequate resources to enhance their capacity for monitoring disinformation and developing situational awareness. Over the past two decades, remarkable advancements in technology have brought about an unprecedented level of digitization, which permeates all aspects of Europe’s society and economy. This digital transformation has opened up new avenues for businesses and significantly enhanced the quality of life for citizens. Moreover, in today's interconnected world, digital platforms and online services have become an integral part of our daily lives, as they facilitate communication and the exchange of information on an unprecedented scale.

However, the rise of social media, e-commerce platforms, smartphones, big data and the increasing use of new digital tools does not only ensure more innovation, but also generates multiple risks and challenges that require a careful assessment of the effects on the economy and society. The lack of adequate regulation of digital platforms poses a range of challenges, encompassing issues such as the rampant spread of disinformation and harmful content, privacy violations and the concentration of market power in the hands of a few dominant Big Tech players and gatekeepers. These challenges have far-reaching negative consequences for both democracy and business competitiveness. In fact, the proliferation of false information has the power to manipulate public opinion, erode trust in institutions, and sow discord within societies. Such divisive narratives can further exacerbate existing societal fault lines, creating an environment of polarisation and mistrust. Therefore, the vulnerability of the digital world therefore requires an effective response from governments and European institutions. Recognizing the urgent need to address these challenges, regulatory frameworks must be established to provide a solid foundation for safeguarding democratic principles, protecting individual privacy, and promoting fair competition. By enacting comprehensive regulations, policymakers can address the dissemination of disinformation and harmful content, enforcing stricter standards of accountability and transparency for digital platforms. This ensures that platforms bear responsibility for the content they host and take proactive measures to counteract false information, while respecting freedom of expression and avoiding undue censorship. Furthermore, measures should be taken to address the concentration of market power in the hands of a few dominant players. The establishment of fair competition policies and robust antitrust regulations can promote a level playing field, fostering innovation, diversity, and equal opportunities for businesses operating in the digital realm. In addition, collaboration between governments, regulatory bodies, and
industry stakeholders is crucial to establish a comprehensive regulatory framework that ensures the responsible and ethical use of digital technologies, while also fostering a thriving and inclusive digital ecosystem.

In this context, the European Union plays a pivotal role in formulating and implementing policies aimed at fostering an online environment that enables businesses and consumers to operate more efficiently, harnessing the full potential of the digital economy.

In fact, as part of the EU Digital Strategy, the European Union has adopted the Digital Services Act as a comprehensive regulatory framework to govern digital services and ensure a safer and more transparent online environment. In particular, the Digital Services Act constitutes the new framework for the regulation of online platforms, introducing a wide range of new rules to strengthen the accountability of these platforms towards their users and society, with regard to issues such as the dissemination of illegal content.

The European Commission believes that this new legislative initiative is set to revolutionize the digital services market in Europe with “two objectives: to promote fundamental rights in digital services and to support technological innovation through the definition of common rules for digital service providers in the European single market”.

As online intermediaries, such as e-commerce platforms, social networks and search engines, have become vital players in the digital economy and an integral part of all aspects of society, updated rules on transparency and accountability in the online space are needed.

The DSA aims to harmonise the regulatory framework for digital services within the EU, updating existing rules that were established more than two decades ago. It seeks to strike a balance between safeguarding user rights, fostering innovation, and holding digital platforms accountable for the content they host and the services they provide.

By establishing clear responsibilities and obligations for digital service providers, the DSA intends to promote user safety, safeguard users fundamental rights, and create a level playing field for all actors in the digital ecosystem.

Moreover, the DSA aims to safeguard users from invasive and unauthorised data collection and marketing profiling, while creating a transparent accountability structure for platforms.

The adoption of the Digital Services Act (DSA) represents a significant step towards a more careful and targeted regulation of the digital services in an ever-evolving digital landscape. While the DSA’s regulatory framework may not address all the challenges associated with digital transformation comprehensively, such as the dissemination of misinformation and the misuse of personal data, it represents a notable effort to establish greater security for users in the virtual dimension.

---

1. In its Communication COM (2020) 67 final, Shaping Europe’s Digital Future, the Commission committed itself to updating the horizontal rules defining the responsibilities and obligations of digital service providers, in particular online platforms.

2. On 15 December 2020, the European Commission, in the context of the EU Digital Strategy, presented a regulatory package on digital services, consisting of two proposals for regulations that are key to improving the functioning of the digital single market: the Digital Services Act (DSA) and the Digital Markets Act (DMA). Specifically, the legislative process of the DSA was concluded on 19 October 2022 with the adoption of the aforementioned regulation, which will become enforceable in 2024. DSA Regulation 2022/2065 was published in the Official Journal of the European Union on 27 October 2022 and will be fully applicable to all entities falling within its scope as of 17 February 2024. However, Article 24(2), (3) and (6), Article 33(3) to (6), Article 37(7), Article 40(13), Article 43 and Chapter IV, Sections 4, 5 and 6 will apply from 16 November 2022. These requirements are only directed at large platforms and large search engines with more than 45 million users per month.


In today’s digital environment, in which novel actors, particularly large technology companies, are “endowed with great economic and, in a broader sense, media powers, the articulated system defined by the Digital Services Act contributes to the reform of the European digital space, ensuring European citizen-users receive a more effective protection both as consumers and as holders of fundamental rights”5. By placing a strong focus on user safety, accountability and enhanced user rights, the DSA aims to create a more transparent and trustworthy online environment. It seeks to strike a balance between fostering innovation and protecting fundamental rights, ultimately aiming to ensure that digital platforms contribute to a more inclusive and resilient digital society. As digital technologies continue to advance and shape our society, it is essential to implement a constantly updated set of regulation in the field, such as the DSA and the related regulatory tools provided by the EU. Through continuous evaluation, adaptation, and effective enforcement, the DSA can foster an environment that balances innovation and user protection, thereby contributing to a thriving and trustworthy digital ecosystem in Europe.

Within this normative framework the Code of Practice on Disinformation6 is a significant component of the European Union’s response to the challenge. It is situated within a wider pattern of EU-level promotion of industry self-regulation of online speech. It joins the ‘Code of Conduct on Countering Illegal Hate Speech (2016)’ and the Commission’s ‘Recommendation on Measures to Effectively Tackle Illegal Content Online (2018)’, each of which complements national legislation restricting these forms of expression to various degrees. All three of these instruments, observes Kuczerawy, are forms of ‘delegated private enforcement’, which tends to be ‘less visible and less obvious’ than direct state intervention. The Code of Practice on Disinformation, however, is unique among these frameworks because it seeks to address lawful speech such as false news articles, conspiracy theories, and hyper-partisan rhetoric. It was first introduced in 2018 as part of the EU’s efforts to address the spread of false information and its potential impact on democratic processes and public trust. The code is a pioneering tool that brought together key industry players to establish self-regulatory standards aimed at combating disinformation. It was the first initiative of its kind, marking a significant milestone in the collective effort to address the challenges posed by false information. By agreeing to this code, relevant stakeholders set a precedent for collaboration and self-regulation in the industry, demonstrating their commitment to combating disinformation and promoting transparency and accountability in the digital realm. The code outlines specific measures to combat the spread of false information, such as improving transparency in online advertising, promoting the visibility of authoritative sources, and developing tools to empower users to make informed choices. It also encourages cooperation and collaboration among different actors, fostering information sharing, and facilitating the development of technological solutions to detect and mitigate disinformation. The 2022 Disinformation Code contains a series of 44 “Commitments,” some of which are further subdivided into “Measures.” When a company becomes a signatory, it submits a subscription document identifying which Commitments (and, more specifically, which Measures) it is signing up for. By adhering to this code, organisations can contribute to maintaining the integrity of information ecosystems, safeguarding democratic processes, and building public trust in the online environment.

Ultimately, the code aims to create a more resilient and trustworthy digital space for all users. Overall, the Code of Practice on Disinformation represents a significant step in the EU’s efforts to address the challenges posed by disinformation, foster transparency, and ensure the responsible behaviour of online platforms in protecting the information ecosystem and democratic values.

Lastly, EDMO (European Digital Media Observatory) and IDMO (Italian Digital Media Observatory) play crucial roles in countering disinformation by providing expertise, research, and coordination efforts.

6The strengthened Code of Practice on Disinformation has been signed and presented on the 16 June 2022 by 34 signatories who have joined the revision process of the 2018 Code. The new Code aims to achieve the objectives of the Commission’s Guidance presented in May 2021, by setting a broader range of commitments and measures to counter online disinformation.
In particular, EDMO is one of the core elements in the Commission’s detailed action plan against disinformation, published in December 2018 and it is conceived to strengthen Europe’s response to disinformation by fostering collaboration and knowledge-sharing among researchers, fact-checkers, and other relevant stakeholders. It supports the development of a European network of researchers and fact-checkers, provides training and resources, and facilitates cross-border cooperation to enhance the detection and debunking of disinformation.

EDMO also conducts research and analysis on disinformation trends and impact, thereby contributing to evidence-based policymaking and the development of effective countermeasures. Both EDMO and IDMO serve as vital hubs for expertise, research, and collaboration in the field of disinformation. By bringing together diverse actors and resources, their efforts are instrumental in promoting a more informed and resilient society in the face of disinformation challenges.

In order to effectively counter the pervasive threat of disinformation, collaboration among policymakers, technology companies, and civil society groups is crucial. By implementing a comprehensive approach that encompasses a range of measures, we can build resilience against disinformation and protect the integrity of public discourse. To effectively address the challenge of disinformation, it is crucial to prioritise the following key areas of intervention:

**a) Enhanced transparency**

Enhancing transparency is a multifaceted endeavour that requires policymakers to take proactive measures and technology companies to embrace responsible practices.

In fact, on one hand, policymakers should encourage and enforce transparency requirements for technology companies, including regulations that mandate disclosure of algorithms.

On the other hand, tech companies should provide users with clear information about the origin, credibility and fact-checking of the content they encounter. This includes information on how content is prioritised, ranked, and recommended to users. By shedding light on these issues, users can better understand how information is presented to them and identify potential biases or manipulation.

Moreover, one should “require platforms to share data around different types of content, by signing on to the Code of Conduct around Platform-to-Researcher data access. Currently, there is some access to data from certain larger platforms, but most of the data is impossible to access. There is an urgent need for an extended access to data for independent researchers, in order to monitor and understand the spread of disinformation. As outlined in the Report of the European Digital Media Observatory’s Working Group on Platform-to-Researcher Data Access published on May 31 2022, a draft code of conduct demonstrates how platform data can be shared safely and ethically with independent researchers. Without this type of access, the scope and scale of the problem remain impossible to assess, and we have no benchmarks by which to measure the effectiveness of any interventions”[1].

In addition to algorithmic transparency, policymakers should also focus on regulating targeted and behavioural advertising in digital services, in order to ensure users’ informed consent.

In fact, “to provide targeted advertising[2], vast masses of users’ data are collected, which involve pervasive surveillance, possibly to the detriment of the concerned data subject, as well as of society as a whole. The processing of such data renders data subjects vulnerable to influence and manipulation as targeted advertising and news shape their online experience, exploit their attention, and direct their actions toward purposes that may not fit their best interests. Moreover, on the providers’ side, targeted advertising promotes concentration, as it gives a key advantage to those companies that given their position—in particular as providers of large platform services— are able to build vast repositories of personal data”[3]. Technology companies should disclose the criteria and mechanisms used to tailor advertisements to users. This transparency will help users understand how their data is being used for perso-

---

nalised advertising and mitigate potential privacy concerns. Policymakers, technology companies, and civil society groups should collaborate to develop industry-wide transparency initiatives. This can involve the establishment of standards and best practices for transparency, the sharing of anonymized data for research purposes, and the creation of platforms or tools that facilitate transparency in online information ecosystems. Collaborative efforts ensure a more comprehensive and effective approach to transparency. By implementing these measures, policymakers and technology companies can work together to establish a transparent information landscape that empowers users, builds trust, and effectively counters the spread of disinformation.

b) Implemented media literacy.

“Media literacy is undoubtedly a crucial tool in the fight against disinformation. A public that is both critically and digitally literate is much more likely to be able to assess the information they encounter online, to identify sources they can trust, and make well-informed decisions as citizens, consumers, and more. Being media literate opens up opportunities to engage more fully and more creatively with the online (and offline) media world. Media literacy education is crucial for adults as well as children. It’s not something that you learn once and then you’re done. It’s about both practical skills, and just as importantly about knowledge and awareness of the digital environment and how it operates. In this sense, it is a life-long process as digital and media environments are constantly evolving”[4]. By incorporating media literacy into educational curricula at various levels, individuals can develop the necessary skills to navigate the vast and often complex digital media landscape. These programs should educate people on how to identify reliable sources, fact-check information, recognize common manipulation tactics, and understand the importance of cross-referencing multiple sources. Additionally, media literacy programs can teach individuals how to identify and critically analyse the underlying motives and biases of media organisations and individuals who produce and distribute content. This empowers users to become more discerning consumers of information and reduces their susceptibility to falling for false or misleading narratives. A media literate person understands the roles and functions of media in society and is able to critically evaluate media content and to interact with media, especially online media, in a mindful way.

A media literate person has a need for media education (which implies that media literacy is a learned competence, that is, it is not an intuitive skill acquired only through the use of media); a need for training in understanding the media (which implies the social role of the media and the need to understand how they are integrated into society); a need for constant questioning of media content (implying a critical attitude of users); and a need for self-expression (implying citizen participation in the communication community, not just their passive role in receiving media messages).

Policymakers should both foster partnerships between technology companies and educational organisations to develop tools and resources that promote media literacy and responsible online behaviour and invest in public awareness campaigns to educate the general public about the risks of disinformation.

Addressing disinformation through digital literacy education is recognized as a key focus in the European Commission’s Digital Education Action Plan (2021-2027)[5]. It proposes legislative actions to mitigate the risks posed by disinformation, such as transparency obligations for platforms, particularly during election periods. The plan also emphasizes the promotion of media literacy, supporting independent fact-checkers, enhancing cooperation with third countries, and protecting media freedom. The plan outlines the Commission’s commitment to developing shared guidelines aimed at equipping teachers and educational staff with the necessary tools to promote digital literacy and effectively combat disinformation through education and training. In fact, “in her political guidelines, President von der Leyen highlighted the need to unlock the potential of digital technologies for learning and teaching and to develop digital skills for all. Education and training are key for personal fulfillment, social cohesion, economic growth and innovation. They are also a critical building block for a fairer and more sustainable Europe. Raising the quality
and inclusiveness of education and training systems and the provision of digital skills for all during the digital and green transitions is of strategic importance for the EU[6].

It is important to recognize that media literacy is not a one-time solution, but an ongoing process that needs to be adapted and updated as new technologies and media consumption habits evolve. Governments, educational institutions, and civil society organizations should collaborate to develop comprehensive and accessible media literacy programs that cater to different age groups, socio-economic backgrounds, and digital literacy levels. Ultimately, implementing media literacy initiatives can equip individuals with the necessary skills and knowledge to identify and combat disinformation, promoting a more informed and resilient society in the face of misinformation and fake news.

c) Implementing Responsible AI

Companies deploying generative AI tools such as ChatGPT and Bard with the potential to generate disinformation should label such content as part of their efforts to combat fake news. OpenAI’s ChatGPT has become the fastest-growing consumer application in history and set off a race among tech companies to bring generative AI products and services to the market.

Concerns are mounting about potential abuse of the technology and the possibility that bad actors and even governments may use it to produce far more disinformation than before[7]. AI has played a complex role in the realm of disinformation. While AI technologies can be used to combat disinformation and improve fact-checking processes, they can also be misused to generate and spread false information at an unprecedented scale. Generative models, such as Deepfakes, can manipulate images, videos, and audio to create convincing but fabricated content. This poses significant challenges for media authenticity and trustworthiness.

AI-powered bots and automated accounts can be employed to disseminate disinformation campaigns, manipulate public opinion, and sow discord in online discussions. Addressing the issue of AI and disinformation requires a multi-faceted approach. On one hand technology companies should prioritise the development and deployment of AI systems that can detect and mitigate disinformation effectively, on the other hand policymakers should enforce policies that promote responsible AI use, while also addressing potential biases and ensuring transparency.

For these reasons, “President von der Leyen announced in her political guidelines for the 2019-2024 Commission A Union that strives for more that the Commission would put forward legislation for a coordinated European approach on the human and ethical implications of AI.

Following that announcement, on 19 February 2020 the Commission published the White Paper on AI - A European approach to excellence and trust that sets out policy options on how to achieve the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of such technology”[8]. In particular, the AI Act legislative proposal aims to strike a balance between fostering innovation and ensuring the protection of fundamental rights and values. On June 14, 2023, the European Parliament and the Council of the European Union endorsed their respective negotiating stances on the Artificial Intelligence Act, which is a proposed legislation on artificial intelligence introduced by the European Commission in April 2021. This approval paves the way for negotiations, referred to as “trilogues,” to finalise the legislation. It is expected that a definitive European regulation will be established by the end of the year, well before the European elections.

The European Commission, with a determination to promote a responsible approach to artificial intelligence systems that enhances the well-being of citizens, has developed its regulatory proposal based on the risks associated with the use of artificial intelligence.

Prohibited will be AI systems that have the potential for intrusive and discriminatory use, posing unacceptable risks
to citizens’ fundamental rights, health, safety, or matters of public interest. This includes applications like manipulating people’s behaviour or vulnerable groups, such as talking toys for children, and social scoring, which involves categorising individuals based on their behaviour or characteristics. AI systems with high risks, such as those employed in critical infrastructures (e.g., power grids, hospitals), decision-making regarding citizens’ lives (e.g., employment or credit assessment), or those significantly impacting the environment, must be regulated through data governance, risk management assessment, technical documentation, and transparency criteria. On the other hand, AI applications used for translation, image recognition, or weather forecasting are considered to have minimal or low risks.

The European Parliament, through its vote on June 14, expands the list of AI systems deemed to have unacceptable risks. This includes biometric categorization systems that utilise sensitive characteristics (e.g., gender, race, ethnicity, citizenship status, religion, or political orientation), predictive policing systems (based on profiling, location, or criminal records), and emotion recognition systems used in law enforcement, border control, workplaces, and schools. The approved text also includes a ban on real-time and post-event remote biometric identification in public spaces, with an exception allowing law enforcement agencies to utilise post-event biometric identification for serious criminal offences upon judicial authorization. However, there might be some steps backward regarding a complete ban on biometric identification during negotiations with the European Council, which had already emphasised in December that AI might be strictly necessary in certain situations, as insisted by many police bodies in member states for crime-fighting activities.

The European Parliament also classifies AI systems that could influence voters and election outcomes, as well as those employed by social media recommendation algorithms or other digital platforms, as high-risk. Moreover, in alignment with an approach that prioritises the needs and interests of humans in their interaction with machines, the approved AI Act allows for the possibility of lodging complaints and receiving explanations for decisions made through high-risk AI systems.

The text approved by the European Parliament for the AI Act includes obligations for providers of generative AI systems and basic AI models. Both types can be considered general-purpose AI as they can perform various tasks and are not limited to a single function. The distinction between them lies in the final output.

Generative AI, like the popular Chat GPT, utilizes neural networks to produce new text, images, or sounds that have never been seen or heard before, akin to what a human can do. To ensure transparency, companies developing generative AI must explicitly indicate in the final output that it was generated by AI. This distinction, for instance, helps distinguish Deepfakes from real images. Such companies must also implement safeguards against generating illegal content and publicly disclose detailed summaries of copyrighted data used to train the algorithm. On the other hand, basic AI models do not “create” but learn from extensive data to perform a wide range of tasks, finding applications in various sectors. Providers of these models must assess and mitigate associated risks related to health, safety, fundamental rights, environment, democracy, and the rule of law.

Additionally, they are required to register their models in the EU Database before entering the market.

In its proposed text for the AI Act, the European Parliament strives to strike a balance between the need to regulate the rapid advancement of technology, particularly considering its impact on citizens’ lives, and the importance of avoiding hindrances to innovation or unfair penalties for smaller companies. Therefore, exemptions have been introduced for research activities and AI components provided under open-source licences, supporting innovation and collaboration within the scientific and technological community. Regulatory sandboxes have also been established to test AI systems and evaluate their ethical, safety, and social impact before their introduction to the market.

Disinformation is one of the major concerns of democratic countries. Detecting and countering disinformation is a
complex challenge due to human, technical, and economic factors.

The rise of human and AI-generated multimedia content complicates matters. Social media platforms and messaging apps are under increased scrutiny from governments and law enforcement worldwide, with concerns about national security and order.

A practical solution may involve offering in-app access to an independent news verification service for users to fact-check content. However, building a reliable service is challenging due to its location-based and diverse nature. Verifying multimedia content, like videos and photos, is also a research challenge.

In the meantime, a multi-pronged strategy combining user education, government regulations, enhanced platform safeguards (like user verification and controlled sharing), and a news verification service supported by reputable agencies can help address the problem.

Over the last three years MediaFutures project funding by the European Commission made commendable efforts in tackling disinformation by conducting mentoring, training and media literacy. With a proactive approach to addressing the evolving challenges of disinformation, particularly concerning AI, demonstrate their commitment to promoting media literacy and responsible AI use. These initiatives play a vital role in empowering individuals and organizations to navigate the complex landscape of information and technology effectively.

[1] EDMO, 10 Recommendations by the Taskforce on Disinformation and the War in Ukraine, June 29, 2022.

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation) defines the practice of profiling as any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements; If the collection and use of data is done in a responsible manner, respecting privacy and with the explicit consent of the individual, profiling is considered a legitimate practice, which helps companies to have a better understanding of the consumer, to undertake targeted marketing actions, to send personalised advertising messages, so as to increase customer satisfaction and loyalty and to record an increase in sales. However, in cases where the creation of a detailed profile of an individual is done without the explicit consent of the individual and through the use of sensitive data, such as sexual orientation or political opinions, profiling is an extremely invasive practice that can cause harm to users. For instance, the data collected may be used to exclude certain commercial offers for certain persons, resulting in discriminatory practices regarding access to certain products or services.


[5] The initiative contributes to the Commission’s priority ‘A Europe fit for the Digital Age’ and to the Next Generation EU.


com/2021/206 final.