SOLARIS

Strengthening democratic engagement through value-based generative adversarial networks

D4.2 SOLARIS
Regulatory innovation framework for infodemic risk mitigation

Lead Author: Andrew McIntyre (UvA)
With contributions from: Tobias Blanke (UvA), Angelo Tumminelli (LUMSA), Yasaman Yousefi (DEXAI), Ledion Krisafi (AIIS), Gjon Rakipi (AIIS), Tommaso Tonello (UU), Lek Hekani (AMI), Maria Dolores Sanchez Galera (UC3M), Paola Saiz (ANSA), Livio Fenga (UNEXE)
Reviewers: Gjon Rakipi (AIIS) & Piercosma Bisconti (DEXAI)

Deliverable nature:	R – Document, report
Dissemination level: (Confidentiality)	PU - Public
Delivery date:	November 2024
Version:	I
Total number of pages:	65
Keywords:	Counter disinformation; generative artificial intelligence (AI); AI-generated media; AI Act; self-regulation; media literacy; cybersecurity; anomaly detection; deepfakes.

Executive summary

This deliverable details work conducted in preparing for the regulatory innovation framework that is aimed at mitigating risks stemming from the circulation and dissemination of AI-generated disinformation and which will be tested during Use Case 2 in M26 (April 2025). Structured into four major sections, the deliverable provides an overview of current policies, strategies and legislation aimed at tackling AI-generated disinformation within Europe with the goal of identifying best practices and determining significant considerations for Use Case 2. Section 1 provides an overview of European-level mitigation strategies including those related to AI governance, cybersecurity, self-regulation of online platforms, and media literacy and education initiatives. Section 2 then explores similar mitigation strategies that are implemented at the national level and attempts to provide a diverse picture of these approaches taken by countries in different European regions, particularly Western Europe, Western Balkans and the Baltic States. Section 3 then briefly explores possible technical detection strategies before presenting the SOLARIS method of anomaly detection which may be implemented into Use Case 2. Finally, Section 4 draws insights from these previous sections in order to propose considerations for Use Case 2 and the ongoing development of the regulatory innovation framework.

Ultimately, the deliverable identifies significant issues facing efforts to tackle AI-generated disinformation, particularly around defining the problem itself and then implementing solutions. Across Europe, AI-generated disinformation is not considered a distinct problem in itself and there is significant divergence and ambiguity around how best to conceptualize the scale and harms of AI-generated disinformation, while often strategies to address the issue are too focused on elections and foreign interference. In attempting to address this ill-defined problem, these diverse and divergent strategies further face challenges in implementation. Notably, the very notion of counter-disinformation is seen to infringe upon freedom of speech and such efforts are often viewed negatively by the public. There are also more practical issues facing implementation in that some States lack the technical infrastructure and government frameworks to effectively combat disinformation, while coordination between different organizations at different levels (local, national, international) poses issues for European States. These issues are further exacerbated by an increasingly fragmented media landscape and fractious online information environment that means users often do not share common news sources thus making it making it more difficult to monitor and combat the spread of disinformation. While there is an increased drive for media literacy education, these initiatives need to continually adapt to technological and political developments and should include more specific education on AI literacy.

To account for the above issues facing efforts to combatting AI-generated disinformation, this deliverable makes specific recommendations for Use Case 2.

Document information

Grant agreement No.	**101094665**	**Acronym**	**SOLARIS**
Full title	**Strengthening democratic engagement through value-based generative adversarial networks**		
Call	**HORIZON-CL2-2022-DEMOCRACY-01**		
Project URL	https://projects.illc.uva.nl/solaris/		
EU project officer	**Ms. I. von Bethlenfalvy**		

Deliverable	**Number**	4.2	**Title**	Regulatory innovation framework for infodemic risk mitigation
Work package	**Number**	4	**Title**	Designing regulatory innovations for infodemic risk mitigation
Task	**Number**	4.3	**Title**	Regulatory innovation framework for infodemic risk mitigation

Date of delivery	**Contractual**	M22	**Actual**	MXX
Status	version I		☒Final version	
Nature	☒Report ☐Demonstrator ☐Other ☐ORDP (Open Research Data Pilot)			
Dissemination level	☒Public ☐Confidential			

Authors (partners)	UvA			
Responsible author	**Name**	Andrew McIntyre		
	Partner	UvA	**E-mail**	a.mcintyre@uva.nl

Summary (for dissemination)	This deliverable will survey existing regulatory options and policies to mitigate GANs risks and contribute to a remediation approach to GANs negative implications to be tested and validated in Use Case 2.
Keywords	Counter disinformation; generative artificial intelligence (AI); AI-generated media; AI Act; self-regulation; media literacy; cybersecurity; anomaly detection; deepfakes.

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**

Acknowledgement


Funded by the European Union

Disclaimer

Table of contents

List of tables

List of figures

Abbreviations and acronyms

ARCOM – French Regulatory Authority for Audiovisual and Digital Communication
AVMSD – Audiovisual Media Services Directive
CENELEC – European Committee for Electrotechnical Standardization
CERT – Serbian Computer Emergency Response Team
CSA – French High Audiovisual Council
DG – Directorate-General
DG CNECT – Directorate-General for Communications Networks, Content and Technology
DSA – Digital Services Act
EDMO – European Digital Media Observatory
EEA – European Economic Area
EEAS – European External Action Service
ENISA – European Union Agency for Cybersecurity
ETSI – European Telecommunications Standards Institute
FITD – Macedonian Fund for Innovation and Technological Development
GCHQ – UK Government Communications Headquarters
GDPR – General Data Protection Regulation
GPAI – General Purpose AI
ICT – Information Communications Technology
IIA – Interinstitutional Agreement on Better Law-making
MStV – German Interstate Media Treaty
NetzDG – German Network Enforcement Act
NIST – US National Institute for Standards in Technology
NLP – Natural Language Processing
Ofcom – UK Office of Communications
RATEL – Serbian Telecommunications Agency
StGB – German Criminal Code
TFEU – Treaty on the Function of the European Union
UNICRI – United Nations Interregional Crime and Justice Research Institute
VLOP – Very Large Online Platform

1 European-level mitigation strategies

This section presents an overview of international legal frameworks and strategies aimed at addressing AI-generated content and online disinformation more broadly within Europe. To begin, the section introduces the European context for governing AI, with particular attention given to the provisions of the AI Act, before then surveying the broader cybersecurity landscape in Europe and initiatives aimed at tackling online disinformation. The section then discusses the EU's approach to platform self-regulation with regards to disinformation and finishes with a discussion on media literacy initiatives within the Europe.

1.1 European and intergovernmental governance context

The EU's campaign to tackle disinformation precedes the emergence of generative AI as the risks of an increasingly digitalized society have become more evident in the last few decades. Since 2015, the EU's work to combat disinformation has involved leveraging on the actions of different stakeholders within the European Commission itself and also in the public and private sectors more broadly.

In the public sector, the European External Action Service (EEAS) and its Strategic Communication Division, with its initial mandate to address Russia's ongoing disinformation campaigns (European Council 2015, 5), are pioneers in the combatting disinformation in the public sector. Additionally, the Directorate-General for Communications Networks, Content and Technology (DG CNECT) is also tasked with tackling foreign information manipulation and interference (EEAS 2021a).

With regards to the private sector, DG CNECT also facilitated with the production of a Code of Practice on Disinformation in 2018 with the intention of encouraging online platforms including Facebook, Google, X, Microsoft and Tik Tok, to play their part in preventing the spread of mis/disinformation. However, the commitments detailed in the Code are self-regulatory and non-binding in nature and, though the Code has since been strengthened in light of the Digital Services Act (DSA), there remain significant concerns that platform self-regulation is not appropriately addressing mis/disinformation. Platform self-regulation will be explored in more detail in section 1.5.

Finally, the European Commission itself has dedicated communication teams that have been analysing, reacting and de-bunking EU-related mis- and disinformation for a long time. Decentralised capacities were brought together for the first time during the Covid-19 pandemic to tackle the scale of threats and the engagement of the Commission has increased as advanced AI technologies have been developed and deployed. This is due to the fact that crucial zones of information and knowledge that support democratic life are increasingly mediated by these systems. The challenges of mis/disinformation are amplified by the power, scale and design of AI technologies as such systems may structure most internet, media and public information domains in the near future (Wihbey 2024). Many of these technologies trigger issues of data governance, such as privacy concerns, data protection, embedded biases, identification and security challenges from the use of data to train generative AI systems, and the negative impacts of generative AI systems.

International cooperation on AI governance is critical for ensuring societal trust in generative AI and for facilitating jurisdictional interoperability in addressing the challenges posed by this technology, including the implications of AI-generated disinformation. Notably, the World Economic Forum has established an AI Governance Alliance that is committed to promoting equal access, shared learning and novel solutions. Further examples of international coordination efforts in drafting AI policy guidance include UNICEF's 2021 Policy Guidance on AI for children and Interpol's 2024 Toolkit for Responsible AI Innovation in Law Enforcement developed in collaboration with the United Nations Interregional Crime and Justice Research Institute (UNICRI). These efforts are grounded on a multistakeholder approach that focuses on knowledge sharing to inform governance cooperation and that embraces a diversity of perspectives from government, civil society, academia, industry and impacted economies. Primarily, governing bodies around the world have issued standards as a method for governing AI. One important example being the British Standards Institution, which launched an AI Standards Hub aimed at helping AI organizations in the UK to understand, develop and benefit from international AI standards. Similarly, the European Telecommunications Standards Institute (ETSI) and the European Committee for Electrotechnical Standardization (CENELEC) have published the European Standardization agenda that includes the adoption of external international standards already available or under development. Meanwhile, the National Institute for Standards in Technology (NIST) in the US has developed an AI Risk Management Framework to support technical standards for trustworthy AI. Despite these efforts to align governance, states around the world have taken different approaches to AI regulation with the US, EU and China diverging on many issues.

In Europe, the Commission's efforts to tackle AI governance have primarily focused on the AI Act which entered into force in August 2024 but will only be fully applicable two years from then. However, there are exceptions: prohibitions will take effect after six months, the governance rules and the obligations for general-purpose AI models will become applicable after one year, and the rules for AI systems that are embedded into regulated products will apply after 36 months. The details of the AI Act will be elaborated on in the next section.

1.2 The AI Act and AI governance

The AI Act is the world's first and only binding regulation on Artificial Intelligence in such a comprehensive way, with an attempt to define an entire sector continuously evolving.

Categories and requirements

In particular, the AI Act outlines regulations for foundation models, defined as general purpose AI models (GPAI) (Article 3, 63). GPAI can be defined as computer models which, through training on a vast amount of data, can be used for a variety of tasks, either individually or as components of a more complex AI system. The AI Act's legal definition, however, is more elaborate and will likely raise objections as it may be interpreted to include a wide range of technologies. This is due to the fact that legal definitions struggle to keep

up with the pace of technological development in an industrial sector that is continuously and rapidly evolving. To keep pace, the AI Act outlines key requirements for all GPAI developers:

1. *Technical documentation* to be kept up-to-date with technical records and shared with downstream AI system providers.
2. *Copyright compliance with EU copyright law* including using advanced technologies (e.g., watermarking).
3. *Transparency in training data* including public releases of detailed summaries of the training data used for their models, following a template provided by the European AI Office (newly established by the European Commission).

The Act primarily targets providers and deployers placing AI systems and GPAI models on the EU market or using them within the EU. There are some exceptions: AI systems used for military, defence or national security purposes by public or private entities are excluded from the Act's scope. Systems developed solely for scientific research and development purposes are also excluded from the Act.

The AI Act introduces a two-tier model to categorize GPAI models as either *basic* or *systemic risk* and that must abide by different regulatory obligations.

Basic GPAI models are subject to mere transparency obligations, consisting of guaranteeing the availability of technical documentation that makes their functioning understandable (also in relation to the data training process) to the European AI Office, as well as to third parties who intend to integrate the model into their AI systems (Article 53, AI Act). This is a reasonable regulation which, in itself, should not constitute a limitation to the development of models. Additionally, GPAI providers operating in the EU must appoint a representative (Article 54, AI Act) and there are further minor regulations for those GPAI models with a free and open license (Article 53, 2). Furthermore, there are separate use cases for GPAI models: (i) where a business or public sector organization uses GPAI in an enterprise setting and (ii) GPAI models that are used by consumers for individual use. Each of these poses distinct risks, some of which may overlap.

The AI Act also introduces the concept of *systemic risk GPAI models* that require additional obligations. The somewhat tautological definition of "systemic risk" provided in the AI Act is as follows: 'a risk that is specific to the high impact capabilities of general-purpose AI models, having a significant impact on the internal market due to its reach, and with actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain' (Article 3, 65). To identify systemic-risk GPAI models, the AI Act indicates that the computational capacity of the model and its size (e.g., very large online platforms (VLOP)) are also criteria (Article 51, 2). A platform or search engine is termed a VLOP if it has more than 40 million monthly users and the AI Act requires reporting for any system with computational capacity of 10^{26} floating point operators (FLOPs), a measure of computational power. However, many new GPAI models display equal levels of performance while remaining under this computational capacity and so do not meet reporting requirements. In addition to the requirements on transparency and copyright protection falling on all GPAI models, provides of

systemic-risk GPAI models are required to constantly assess and mitigate the risks they pose and to ensure cybersecurity protection. That includes keeping track of, documenting and reporting serious incidents.

Wachter (2024) explains how this regulatory approach is disappointing. The two-tier model is quite unconvincing because many "systemic risks" occur in all GPAI models, regardless of their size or computation. This means that misinformation, data protection issues, explainability problems and harmful outcomes occur in smaller and less "capable" systems. These critical aspects have also raised by Weidinger et al (2022).

In addition to this two-tier model, the European Commission has also identified prohibited AI systems and has set technical standards and best practices for AI system design to prevent manipulative techniques (Article 5).

Supranational authorities for AI governance

The AI Act mandates a comprehensive governance framework that establishes the European AI Office as the main supranational authority and that oversees AI advancements, liaise with the scientific community, and play a pivotal in investigations, testing, and enforcement, all with a global perspective. This framework is based with the DG CNECT and the proposed governance structure involves both national and supranational bodies.

The AI Office (Article 64, AI Act) plays a pivotal role in applying and enforcing regulations concerning GPAI, focusing on standardization efforts to harmonize tools, methodologies, and criteria for evaluating systemic risks associated with GPAI across supranational and national levels. It also monitors GPAIs for adherence to standards and potential new risks, supports investigations into GPAI violations, and assists in developing delegated acts and regulatory sandboxes for all AI systems under the AI Act. It provides administrative support to other bodies: the European AI Board, the Advisory Forum and the Scientific Panel. It further consults and cooperates with stakeholders, with other relevant Directorates-General (DGs) and services of the Commission, as well as fostering international cooperation.

The European AI Board (Article 65, AI Act), instead, includes representatives from each Member State, with the AI Office and the European Data Protection Supervisor participating as observers. The Board coordinates national competent authorities and harmonizes administrative authorities and administrative practices. They further issue recommendations and opinions upon request of the Commission). Finally, the Board supports the establishment and operation of regulatory sandboxes and gather feedback on GPAI-related alerts.

The Advisory Form (Article 76, AI Act) is a group of stakeholders appointed by the Commission that provides technical expertise and prepare opinions and recommendations upon request of the Board and the Commission. The Forum also establishes sub-groups for examining specific questions and prepares annual reports on activities.

The Scientific Panel (Article 68, AI Act) involves independent experts selected by the Commission to support enforcement of AI regulation (especially for GPAI models with the goals of improving GPAI performance), provide advice on the classification of AI models with systemic risk, alert the AI Office of

systemic risks, develop evaluation tools and methodologies for GPAIs, and support market surveillance authorities and cross-border activities.

National authorities for AI governance

Within the AI Act's governance framework, national authorities play a key role. The responsibilities assigned to this governance structure are still broadly defined, with the expectation that their precise implementation will evolve based on practical experience.

Notifying authorities (Articles 28-29, AI Act) are designated or established by Member States to process applications for notification from conformity assessment bodies (CABs), monitor CABs, cooperate with authorities from other Member States, ensure no conflicts of interest with CABs, conflicts of interest prevention and assessment of impartiality.

Notified bodies (Articles 29-38, AI Act) are third-party conformity assessment bodies (with legal personality) that verify the conformity of high-risk AI systems, issue certifications, manage and document subcontracting arrangements and issue periodic assessment activities (audits). They also participate in coordination activities and European standardization.

Market surveillance authorities (Articles 70-72, AI Act) are entities designated or established by Member States as single points of contact. They issue non-compliance investigations and corrections for high-risk AI systems, real-world testing oversight and serious incident report management. They guide and abvice on the implementation of the regulation, particularly to SMEs and start-ups, they also provide consumer protection and fair competition support.

The AI Act is not, however, the only regulatory instrument that is set to govern AI. It is complemented by cross-legal harmonized standards by two liability frameworks: the Product Liability Directive (PLD) and the Artificial Intelligence Liability Directive (AILD). More broadly, the AI Act relates to the European cybersecurity landscape.

1.3 European cybersecurity landscape

As digital technologies have become more deeply integrated into society in recent years, cyber threats including hacking and social engineering efforts have grown in scale, complexity and sophistication. The EU has responded to these growing threats with comprehensive cybersecurity strategies aimed at protecting critical infrastructure, ensuring the privacy of personal data, promoting trust in digital technologies, and combatting political influence operations and disinformation campaigns. This section provides an overview of the core cybersecurity strategies at a European-level focusing on regulations, directives and overarching frameworks that are crucial in addressing these challenges, particularly in mitigating against the risks of disinformation. While these strategies are often more focused on improving technical security of products and infrastructure, protecting data privacy, and locating harmful content online, there is a convergence between disinformation and cybersecurity that is becoming increasingly relevant for combatting political manipulation (EU Disinfo Lab 2021). Fundamentally, online disinformation and cyber threats operate on the same strategic terrain and

spread via the same information communication channels such as social media platforms, networking infrastructure, and routing services. While significant national and international cybersecurity strategies have been introduced and organizations have become far more resilient to traditional cyberattacks, there is a lack of effective strategies for combatting disinformation as well as a lack of preparedness and literacy education. As such, cybercriminals and bad actors are increasingly using disinformation tactics (e.g., defamation, extortion) and furthermore integrating disinformation into traditional cyberattacks forming highly-effective hybrid threats (e.g., using convincing disinformation content to deliver malware). Finally, the targets and victims of cyberattacks and disinformation campaign are often similar (e.g., influential individuals, government institutions, high-profile private businesses). Therefore, while many of these European-level cybersecurity strategies do not explicitly apply to disinformation and AI-generated disinformation, they will continue to play a key role in ensuring the EU's resilience against political manipulation and may offer insights and solutions for combatting this problem.

Cyber Resilience Act

Proposed in 2022, the Cyber Resilience Act has the goal of ensuring that products with digital elements, such as hardware and software, meet high cybersecurity standards (European Commission 2022d). This legislation comes at a time when the EU is increasingly exposed to large-scale cyberattacks, including ransomware and supply chain attacks, which have the potential to disrupt entire sectors. The Act mandates manufacturers to build security measures into their products from the design phase and to maintain those standards throughout the lifecycle of the product, including offering timely security updates. The Cyber Resilience Act addresses a significant gap in cybersecurity regulations by introducing horizontal cybersecurity requirements for products across all sectors. It aims to protect users from vulnerabilities that could be exploited by malicious actors and to prevent the widespread impact of cybersecurity incidents. Through this Act, the EU aims to bolster the security of its digital market and ensure that businesses and consumers alike can operate securely within an interconnected digital environment. Notably, the Cyber Resilience Act emphasizes transparency and accountability in that manufacturers are required to disclose security features and vulnerabilities, allowing users to make informed choices about the digital products they purchase. This transparency is crucial in fostering trust in the digital economy, which is essential for the successful functioning of the European single market.

EU Cybersecurity Act

Adopted in 2019, the EU Cybersecurity Act is a cornerstone regulation that strengthens the role of the European Union Agency for Cybersecurity (ENISA) (European Commission 2019b). The Act enhances ENISA's mandate, making it a permanent entity responsible for overseeing the EU's cybersecurity landscape. Additionally, the Cybersecurity Act introduces a certification framework for ICT (information and communication technology) products, services, and processes. The cybersecurity certification framework is a vital component of the EU's efforts to harmonize cybersecurity standards across Member States. By offering a standardized certification process, the Act aims to reduce fragmentation in the cybersecurity market, enabling

easier cross-border transactions and fostering a more secure digital single market. Furthermore, this certification ensures that ICT products meet stringent security requirements, reducing the risk of cyberattacks and boosting consumer confidence in digital technologies. The EU Cybersecurity Act also highlights the need for reducing dependency on non-EU cybersecurity products and fostering innovation within the EU. By By promoting homegrown cybersecurity solutions, the Act aims to enhance the EU's digital sovereignty, ensuring that critical infrastructure and services are protected by solutions that are developed and governed within the EU.

NIS2 Directive

The NIS2 Directive is an updated version of the original Network and Information Systems Directive which was the first EU-wide cybersecurity law introduced in 2016. The updated directive, adopted in 2022, expands the scope of cybersecurity regulations to cover more sectors and industries, ensuring a high level of cybersecurity across the EU (European Commission 2022c). The directive is particularly important for critical infrastructure sectors, such as energy, healthcare, transport, and finance, which are highly dependent on secure network and information systems. The NIS2 Directive strengthens cybersecurity requirements by imposing more stringent risk management and incident reporting obligations. It also introduces a framework for coordinated cyber crisis management and establishes penalties for non-compliance. One of the key objectives of the NIS2 Directive is to enhance cross-border cooperation and information sharing among Member States. Cyber threats are often not confined to national borders, and a coordinated response at the EU level is essential for addressing incidents that may affect multiple countries or industries simultaneously. The directive also provides for stronger regulatory oversight, ensuring that Member States implement consistent and robust cybersecurity practices across the EU.

General Data Protection Regulation

The General Data Protection Regulation (GDPR), implemented in 2018, is one of the most influential pieces of legislation in the realm of data protection and cybersecurity (European Commission 2016b). Although primarily focused on personal data protection, the GDPR has significant implications for cybersecurity. The regulation sets strict standards for the collection, processing, and storage of personal data, requiring organizations to implement appropriate security measures to protect data from unauthorized access, loss, or theft. The GDPR has been instrumental in shaping the global conversation around data privacy and security. Organizations must ensure that they have robust cybersecurity practices in place to comply with GDPR requirements, including encryption, access control, and incident response procedures. In the event of a data breach, organizations are required to notify relevant authorities within 72 hours, providing transparency and accountability in the face of cybersecurity incidents. By imposing substantial fines for non-compliance, the GDPR has incentivized organizations to take cybersecurity seriously. The regulation has also helped to align data protection practices across the EU, ensuring a high level of security for personal data and contributing to overall cybersecurity resilience.

Digital Services Act

Introduced in 2022, the Digital Services Act (DSA) is a comprehensive framework aimed at regulating online platforms and services. In the context of cybersecurity, the DSA is particularly relevant as it addresses the responsibilities of online platforms in mitigating risks such as illegal content, disinformation, and cyberattacks (European Commission 2022b). The act introduces new obligations for online platforms to enhance transparency, accountability, and user protection. One of the critical elements of the DSA is its focus on large online platforms, which have a significant influence on the spread of information and content. Article 33 of the Act focuses on "online platforms and online search engines which have a number of average monthly active recipients of the service in the Union equal to or higher than 45 million, and which are designated as very large online platforms or very large online search engines" (Ibid.). These platforms are now required to implement proactive measures to prevent the dissemination of harmful content and improve their response to cybersecurity incidents. Additionally, the DSA mandates greater transparency in content moderation processes, helping to build trust in digital platforms and services. By establishing a safer online environment, the DSA complements the EU's broader cybersecurity objectives. It seeks to reduce the risks associated with online disinformation campaigns, which can undermine trust in democratic institutions and exacerbate the impact of cyber threats.

EU Cybersecurity Strategy for the Digital Decade

Published in 2020, the EU Cybersecurity Strategy for the Digital Decade outlines the EU's long-term vision for building a secure and resilient digital environment (European Commission 2020b). This strategy emphasizes the need for cybersecurity to be a fundamental pillar of Europe's digital transformation, with a focus on protecting critical infrastructure, enhancing public-private partnerships, and promoting international cooperation. The strategy addresses the growing cybersecurity risks associated with the rapid digitization of industries and services. It highlights the importance of investing in cutting-edge cybersecurity technologies, including AI systems and quantum computing, to stay ahead of evolving cyber threats. Furthermore, the strategy advocates for stronger regulatory frameworks to ensure that cybersecurity becomes an integral part of digital innovation and economic growth. An essential aspect of the strategy is its focus on global cooperation. Cyber threats are a global issue, and the EU aims to take a leadership role in promoting international standards for cybersecurity. The strategy also emphasizes the need for closer collaboration with international organizations, such as NATO and the United Nations, to enhance global cybersecurity resilience.

European Digital Identity Framework

The European Digital Identity Framework, set to be fully operational by 2024, is another significant initiative in the EU's cybersecurity landscape (European Commission 2024). This framework aims to provide EU citizens with a secure and interoperable digital identity that can be used across borders for both public and private services. The digital identity will offer a trusted means for individuals to authenticate themselves online, reducing the risks of identity theft, fraud, and unauthorized access to personal data. The framework supports the EU's broader cybersecurity goals by ensuring that individuals have greater control over their

personal data while accessing online services. It also facilitates secure cross-border transactions, contributing to the EU's ambition of creating a fully integrated digital single market.

European Strategy on Countering Hybrid Threats

Hybrid threats, which combine cyberattacks with disinformation campaigns, have become a significant concern for European security (European Commission 2016a). The EU Strategy on Countering Hybrid Threats, introduced in 2016, provides a framework for addressing these multifaceted challenges. The strategy emphasizes the need for a coordinated response that leverages both national and EU-level resources to protect critical infrastructure and democratic institutions. The strategy highlights the importance of enhancing situational awareness, information sharing, and joint response capabilities to counter hybrid threats. It also underscores the need for closer collaboration with international partners to address the global nature of hybrid threats, which often target both military and civilian infrastructure.

Generative AI in the cybersecurity landscape

The EU has fostered a robust and comprehensive cybersecurity landscape that addresses the multi-faceted challenges posed by digital technologies including a solid legal foundation that ensures that businesses, individuals, and governments can operate securely in the digital environment. Furthermore, these efforts have been bolstered by key strategic initiatives aiming to build a secure, resilient and integrated digital single market with the aim of building trust in the digital economy. However, disinformation is repeatedly overlooked as a cybersecurity issue and there is a lack of specific strategies to address this challenge. With the arrival of generative AI, cyberattacks and disinformation campaigns may become much more effective, sophisticated and complex requiring cybersecurity strategies to keep pace and adapt in order to respond.

1.4 Platform self-regulation

As online platforms have become increasingly politicized spaces and the predominant channel by which mis/disinformation is spread through communities, there has been increasing tension between the preservation of users' freedom of speech and the need to moderate harmful and/or illegal content on these platforms. In Europe, this has been characterised as a problem of outsourcing in that governments have requested that large technology companies operating these platforms take appropriate steps to ensure a safe online environment for all citizens. This section analyses the notion of self-regulation for online platforms operating within the European Economic Area (EEA) and specifically focuses on the steps taken by these platforms to address the spread of mis/disinformation, particularly in light of recent technological developments with generative AI. The section begins with an overview of the legal tools EU institutions use to tackle policy challenges by delegating executive powers and then delves into the methods of combatting mis/disinformation in more detail. Primarily, these methods include codes of practice (self-regulatory frameworks shaping how platforms/search engines may autonomously tackle mis/disinformation) and articles from regulations (specific sections of legal binding documents referring to self- and co-regulation practices).

Overview of European soft law tools

Since the 1984 "Guidelines for Co-operation", the European Commission has been trying to approach and innovate supranational governance and in specific instances to delegate executive powers to private stakeholders represented within CEN-CENELEC and other regulatory bodies. More specifically, the Commission was interested in entrusting such private international non-profit organizations with setting out European standards, also called "harmonised standards" as to implement its Directives. While Directives "limit themselves to elaborating 'essential requirements' with which products must conform…the task of drawing up the technical specifications is left to the European standardisation bodies" (Schepel & Harm, 2005). By relying on bodies of "technically competent experts", the outsourcing procedure had the goal of making the legislative process swifter, more pragmatic and more representative of the interests of different categories affected by new legislation (Enhancing the Implementation of the New Approach Directives, 2003). Even though Directives are broadly phrased, the details elaborated by European Standards Bodies are binding in nature. As such, this outsourcing has since been characterized as a form of privatisation in that the Commission is seen to be delegating the work of public interest to private organizations and has thus been criticized for a lack of appropriate supervision by the Commission. Furthermore, the slow drafting and implementation of good governance practices has led to notoriously overly complicated and untimely regulation by the European Standards Bodies (Schepel & Harm 2005). In response to these issues, the Commission has turned to *soft law* measures which, instead, are focused on "promoting dialogue, inventing co-operative linkage institutions, encouraging the flow of information and mutual learning and adaptation between the public and private spheres" (Schepel & harm 2005).

In the 2001 White Paper on European Governance, the Commission explicated its interest in promoting a more frequent use of different policy tools with respect to the ones traditionally envisaged by EU law, thus signalling the Commission's intention to incorporate non-binding tools into the EU normative framework and legislative mechanism (Senden et al. 2015; European Commission 2001). The inclusion of soft law instruments such as self- and co-regulatory mechanisms was meant to further enhance the quality, speed and flexibility of the legislative process, while partially overcoming privatisation issues. By 2003, the Interinstitutional Agreement on Better Law-making (IIA) officially recognised the possibility of using alternative regulatory mechanisms in cases where legal instruments were not specifically required by the Treat on the Function of the European Union (TFEU) (Senden et al. 2005). In the IIA, self-regulation is defined as "the possibility for economic operators, the social partners, non-governmental organisations or associations to adopt amongst themselves and for themselves common guidelines at European level (particularly codes of practice or sectoral agreements)" (IIA 2003). The non-binding nature of self-regulation stresses the aforementioned interest in encouraging dialogue and "mutual learning and adaptation between the public and private spheres" (Schepel & Harm 2005).

While broadly, soft law tools such as self- and co-regulation all involve gathering together private stakeholders and asking them to take part in drafting sets of principles they are not legally bound to respect, there are some differences. While self-regulation entails private stakeholders (e.g., firms, consumer advocacy groups) drafting guidelines following an invitation to do so by the European Commission, co-regulation requires the Commission as the governing body to set "default requirements and retains general oversight authority to approve

and enforce these guidelines" (Selinger et al. 2018). There are, of course, incentives for stakeholders to take part in these consultations. As a bottom-up approach, self-regulation empowers private agents to contribute to the drafting of these principles and in doing so they can represent their specific concerns and ensure such principles reflect market realities. Furthermore, taking part in these consultations may allow private stakeholders to inform future binding EU regulation (De Witte 2023). Having clarified the role for self- and co-regulation inside the European Common Market, the next sections present instances of their application that are specifically relevant to the issue of AI-generated disinformation.

Codes of practice

In the context of self-regulation to address the spread of mis/disinformation on online platforms, the European Commission has pursued tow main sets of principles: (i) the *2018 Code of Practice on Disinformation* and (ii) the *2022 Strengthened Code of Practice on Disinformation*. The peculiarity of these Codes of Practice stems from the fact that they are in practice, non-binding legal instruments. Hence, companies may decide whether to undertake a commitment towards the European Commission and to what extent participate in such agreements. Moreover, both of the Codes were created following the Commission's willingness to involve industry partners in drafting shared guidelines to tackle online disinformation (European Commission 2018; European Commission 2021).

2018 Code of Practice on Disinformation

The purpose of the *2018 Code of Practice on Disinformation* is to "identify the actions that Signatories could put in place in order to address the challenges related to 'Disinformation'". Most notably, the Code provides for Signatories to: implement policies to stop monetisation of misleading content; define and publish rules on the misuse of automated systems; invest in tools to help users identify false news and develop trust indicators; prioritise reliable information in search and feed algorithms; collaborate with stakeholders to improve digital literacy and critical thinking; provide annual reports on efforts to counter disinformation, including transparency, public awareness, and training initiatives; and, finally, use an independent third party to review and assess progress on these commitments. However, as it is recognized that Signatories "operate differently, with different purposes, technologies and audiences, the Code allows for different approaches to accomplishing the spirit of the provisions" (European Commission, 2018). As such, the Signatories are free to decide which objectives they are committed to and to what extent they pursue these objectives. Due to its non-binding and flexible nature, the Code may be understood as a non-compulsory tool intended to attract as many stakeholders as possible and to further monitor and study the effects of self-implemented policies with the objective of drafting more precise and effective regulatory frameworks in the future. The Code improved the Signatories' accountability by setting broad policy goals, defining necessary actions, and allowing public disclosure. These insights later informed the European Commission's policy initiatives, including the European Democracy Action Plan and the Digital Services Act, which established overarching rules for all information society services (European Commission 2020).

While the Code broadly sought to address a diverse set of challenges stemming from online mis/disinformation, its very approach was later criticised in 2020 in an internal Commission assessment (Staff Working Document 2019). This assessment highlighted that there were several severe shortcomings in the Code

in that it lacked shared definitions, clear procedures and precise commitments, and did not enable better monitoring by European institutions. The assessment also encouraged additional efforts to be made to include other critical stakeholders, especially from the advertising sector. Furthermore, the assessment highlighted that it was difficult to assess the impact of Signatories' actions over time due to the reliance on platform willingness to share such information. A more structured cooperation between platforms and the research community had already been advocated in 2019 by the French Mission Report, which highlighted issues of transparency and legitimacy by Meta (then Facebook), by far the largest private online stakeholder with around 2.5 billion monthly users at the time (French Mission Report 2019). In 2021, the European Commission further elaborated on the 2020 assessment and sought to strengthen the Code (European Commission 2021) acknowledging the efforts of the 2020 Joint Communication on Tackling Covid-19 Disinformation which encouraged Signatories of the Code to report on their actions to tackle fake news regarding the pandemic (European Commission 2020). Recognising novel disinformation threats and the renewed challenges for European democracies, the Commission underlined the importance of transforming the Code into a living instrument by setting up a "permanent mechanism for its regular adaptation" (European Commission 2021). This entails actively modifying the content of the Code with the goal of tackling new disinformation challenges as they arise.

These assessments of the Code led to the drafting and adoption of the 2022 Strengthened Code of Practice on Disinformation in order to address these limitations and shortcomings.

2022 Strengthened Code of Practice on Disinformation

The 2022 Code represents a more structured and exhaustive framework for the private companies committing to tackling online disinformation on their platforms and is, to some extent, a reaction to the Covid-19 pandemic. Since the beginning of the pandemic, EU institutions sought to adopt soft-law acts to tackle the spread of disinformation related to Covid-19 as 'due to their informal status, soft law measures can be adopted more quickly than formal laws' enabling a rapid response to the emergency (De Witte 2023). Intended as an evolution and improvement upon the previous 2018 Code, the Strengthened Code promotes the following measures: demonetisation of disinformation content; tackling advertising containing disinformation; adopting shared definitions of political and issue advertising/impermissible manipulative behaviour; transparency obligations for AI systems (e.g. deepfakes); enhancing media literacy; flagging harmful false and/or misleading content; disclosure of and access to signatories' data for research on disinformation; establishing of a permanent taskforce to monitor and adapt the Code to address future challenges (European Commission 2022f). The potential commitments specified in the Strengthened Code also differ depending on the size of the platform, following the DSA's definition of VLOPs. As such, smaller platforms with less than 45 million users per month may scale down the measures they undertake to pursue the Strengthened Code's commitments (Regulation 2022/2065).

A notable development in the Strengthened Code is the introduction of qualitative reporting elements and service-level indicators/quantitative indicators which aim to track the actions taken by Signatories and to closely measure their application. This development seeks to address issues of transparency identified by the French Mission Report and by the Commission's own assessment of the 2018 Code. In spite of this, online companies have continued to provide unsatisfactory information while voluntarily complying with the standards set by

themselves. Notably, there have been many instances in which online platforms have only shared partial information, provided inaccurate "methodologies for calculating quantitative data", not sufficiently enforced "bans on political advertisements and actions against fake engagement and inauthentic behaviour", and provided insufficient data at the Member State-level (Mündges & Park 2024).

Given the fairly recent implementation of the 2022 Strengthened Code, scientific literature monitoring its implementation and impacts is still scarce. Furthermore, there is a lack of data on disinformation in Europe generally and on the actual effects of measures that online companies have taken to tackle such issues. As such, the qualitative reporting elements and quantitative indicators introduced by the Strengthened Code have become an imperfect or incomplete tool for monitoring progress in tackling disinformation (Nenadic et al. 2023). Some have argued that the Strengthened Code represents a step towards co-regulation (Parcu & Brogi 2021), while others such as Mündges and Park (2024) speculated that the 2022 Code of Practice could have been turned into a Code of Conduct with binding value following the introduction of the DSA. Even though this transformation did not occur, the Commission has shown its willingness to provide more systematic guidance to such instances of self-regulation.

With regards to generative AI specifically, from 2023 Signatories of the Strengthened Code started to voluntarily report on their strategies for addressing AI-generated media and deepfakes. This measure would eventually be made compulsory in the AI Act. The Code also emphasizes the demonetization of disinformation, aiming to reduce the financial incentives for creating misleading content, thereby discouraging its dissemination. Cooperation with independent fact-checkers is also a vital aspect of the Code, as it enhances the identification of false information and provides users with reliable tools for verifying claims. Moreover, the transparency requirements of the Code could aid in recognizing misleading AI-generated media, particularly those designed to mislead voters or consumers.

Articles from regulations

Beyond codes of practice, it is also possible to identify soft law provisions within binding legislation. That is, while hard law documents (e.g., AI Act, DSA) set out compulsory requirements, they may also contain Articles that impose non-binding obligations. More specifically, in both the DSA and the AI Act several articles encourage the further implementation of Codes of Conduct hinting at EU institutions' desire to retain and widen the use of self-regulation tools. However, considering the need for more substantial and univocal improvements in tackling online disinformation, it is necessary to ask whether a co-regulatory approach, in which there is closer and more regular supervision by public authorities, is necessary.

The DSA focuses on providing EEA-wide regulation to tackle "diligence requirements for providers of intermediary services as regards the way they should tackle illegal content, online disinformation or other societal risks" (Regulation 2022/2065). In this respect, Article 7 and Article 8 of the DSA state that companies may undertake voluntary own-investigation initiatives and, while acting good faith (i.e., they do not fail to act on or report the presence of illegal/harmful content), they may not be considered responsible for the kind of data they are storing or transmitting. Article 19 and Article 29 of the DSA further specify different obligations and exemptions from any binding norms for small or micro-firms that would not be able to make the necessary

investment to fulfil the goals of the DSA. Nonetheless, these companies are expected to act based on the means and resources available to them as invoked by the 2022 Strengthened Code. Finally, Article 44 and Article 45 of the DSA encourage the drafting of new Codes as complementary measures to more swiftly achieve the objectives of the Act.

Notably, the above exemptions for small or micro-firms represent an important limitation of the DSA as harmful online content may still circulate widely via online platforms with less than 45 million monthly users. Hence, the DSA allows for a high-degree of self-regulation for smaller companies and these companies are theoretically free to "propagate disinformation or enable discrimination" without breaching the provisions of the Act (Nannini et al. 2024).

Since the introduction of the DSA in 2022, generative AI technologies have developed significantly and rapidly such that the threat of AI-driven disinformation has only increased. As such, the DSA only partially addresses the issue of AI governance in relation to online disinformation. With regards to AI-driven disinformation, the AI Act also has significant shortcomings. While Article 69 of the AI Act encourages the drafting of new Codes to tackle future challenges stemming from AI systems, disinformation is only briefly mentioned in the preambulatory clauses of the Act itself. It is acknowledged that AI systems represent new challenges to online users but this issue is not further elaborated on in the main provisions of the Act (Regulation 2024/1689). Neither the DSA nor the AI Act would seem to sufficiently address the specific issue of AI-generated disinformation. In September 2024, the Commission launched a call for consultations on a new Code of Practice for providers of GPAI models; the kind of self-regulatory measure encouraged by both the AI Act and DSA.

Limitations of self-regulation

Self-regulation and co-regulation practices are becoming more and more common within the EU legislative process and their ability to evaluate and adjourn exiting norms or to fill normative gaps represents an important source of information for EU institutions. The experimental and non-binding design that codes of conduct, recommendations and guidelines exhibit represents their very strength: their ability to attract and retain large numbers of stakeholders informs legislators of best practices and of potential complications when shaping legally binding documents. However, the effectiveness of self-regulation measures has been criticized and these measures have largely failed to translate into detailed regulations. Notably, EU regulations have only been able to address the compliance of VLOPs and the legislative process is still not able to keep pace with real-world developments. Despite these criticisms, the EU continues to rely upon self-regulatory tools. While Codes of Practice are in place, there remain questions around how to ensure online companies adhere to these principles and whether or not such Codes should be bound to EU regulations. On the other hand, there is increasing concern around the effectiveness of such regulations and the role of both hard and soft law methods as highlighted by Mario Draghi (Draghi 2024). As Draghi highlights, stringent regulations may stifle competition, investment and industrial innovation in Europe, particularly with regards to generative AI, such that the EU may fail to keep pace with economic rivals such as the US and China.

1.5 Media literacy initiatives in Europe

As digital media has begun to permeate every aspect of life, media literacy has become an essential skill for European citizens. Recognizing this, the European Commission and various other pan-European organizations have introduced a range of initiatives aimed at enhancing media literacy across the continent. Indeed, media literacy plays a crucial role in empowering citizens and fostering critical thinking, enabling individuals to critically evaluate media content and become more discerning consumers. This enhanced capacity is essential for identifying credible sources and effectively combating mis/disinformation, particularly in light of the increasing prevalence of disinformation campaigns across Europe. Moreover, media literacy enhances the public's understanding of media production and tackles the inherent biases that often accompany it. This is vital for building resilience against mis/disinformation, especially during pivotal events such as elections and public health crises, where the spread of false information can have significant consequences (European Commission 2022e). Ultimately, media literacy serves as a cornerstone for informed and responsible digital citizenship, equipping individuals to navigate and contribute positively to the digital environment.

As the emergence of generative AI has exacerbated the issues of online mis/disinformation, media literacy initiatives have become more vital than ever. This section examines the key media education strategies implemented in Europe, focusing on initiatives designed to combat mis/disinformation and the specific challenges posed by AI-generated media. It will also explore the strengths and limitations of these initiatives, highlighting their implications for the future of media literacy education.

To begin, the European Commission emphasizes the importance of digital literacy as a fundamental component for active participation in contemporary society, advocating for media literacy programs that aim to bridge the digital divide. By equipping diverse populations with the skills necessary to engage responsibly with digital content, these initiatives promote inclusivity and ensure that all individuals can navigate the complexities of the digital world (European Commission 2022a). The European Commission has made significant strides in enhancing media literacy through a variety of initiatives designed to equip citizens with the skills necessary to critically engage with digital content. The Digital Education Action Plan (2021-2027) is a cornerstone of these efforts, emphasizing the integration of digital literacy into educational systems across Europe. This plan not only aims to develop critical thinking skills among students but also supports educators by providing resources and fostering collaboration between member states to share best practices in media education (European Commission 2021a). The initiative underscores the importance of a coordinated approach to media literacy, ensuring that all Member States can benefit from shared knowledge and resources.

In addition to the Digital Education Action Plan, the European Media Literacy Week, launched in 2019, serves as a key platform for raising awareness about the importance of media literacy. This annual event brings together stakeholders from various sectors, including policymakers, educators, and civil society organizations, to discuss the challenges and opportunities associated with media literacy education. The event also provides a venue for showcasing successful initiatives and strategies, fostering a sense of shared purpose across the continent (European Commission 2019a). Complementing these efforts is the Audiovisual Media Services Directive (AVMSD), revised in 2018, which mandates EU member states to actively promote media

literacy, particularly in the context of digital and online platforms. This directive highlights the crucial role of media literacy in protecting users from harmful content and misinformation, providing a regulatory framework that supports the broader goals of media education in Europe (European Parliament 2018a).

These initiatives offer numerous benefits, including fostering cross-border collaboration and emphasizing the importance of formal education in developing critical thinking skills. They also support ongoing professional development for educators, ensuring that they are equipped to teach media literacy effectively. Additionally, many media literacy initiatives incorporate cutting-edge technologies such as generative AI and virtual reality (VR), which not only improve learning experiences but also prepare individuals for the continuously evolving media landscape. These innovative approaches engage learners in immersive ways, fostering deeper comprehension and retention of critical media skills (European Commission 2018c). However, there are challenges associated with these initiatives, such as the varying levels of implementation across Member States, which can lead to disparities in media literacy. Additionally, reaching older adults and marginalized communities who may be less engaged with formal education systems remains a significant hurdle.

In response to the rising tide of fake news and disinformation, particularly during the COVID-19 pandemic, the European Commission and other bodies have intensified their efforts to combat these threats. The European Digital Media Observatory (EDMO), established in 2020, plays a crucial role in this regard. EDMO serves as a collaborative hub for fact-checkers, researchers, and other stakeholders working to counter disinformation. It supports the development of tools and resources to detect and mitigate false information, with an increasing focus on AI-generated disinformation (EDMO 2020).

While these initiatives have made substantial progress in fostering collaboration and raising public awareness about the dangers of disinformation, they also face significant challenges. The voluntary nature of the Code of Practice, for instance, limits its enforceability, raising concerns about its long-term effectiveness. Additionally, the rapid development of AI technologies used to generate disinformation presents ongoing challenges for detection and verification of misleading news, as current tools may struggle to keep pace with these advancements. As AI-generated media becomes more sophisticated, the need for continuous innovation in both media literacy education and technological tools becomes increasingly apparent.

AI-generated media, including deepfakes and synthetic news, pose a significant challenge to media literacy initiatives. These technologies can create highly realistic and convincing content, making it increasingly difficult for users to distinguish between real and fake information. In response, media literacy initiatives are placing greater emphasis on educating users about the existence and potential dangers of AI-generated media. This includes developing critical thinking skills and providing tools to help users verify the authenticity of content they encounter online. There is also a growing focus on technological solutions, with collaboration between technology companies and educational institutions being essential for the development of tools capable of detecting AI-generated content.

Despite these efforts, several challenges remain. While raising awareness about AI-generated content is crucial, detection tools may not be accessible nor understandable for all users, particularly those with lower levels of digital literacy. Additionally, the constantly evolving nature of AI-generated media requires

continuous updates to educational content and tools. There are also ethical concerns surrounding privacy and surveillance in the development and deployment of detection technologies, which need to be carefully considered.

Media literacy education in Europe is a vital component of the region's strategy to combat mis/disinformation, particularly in an increasingly digital world. While significant progress has been made through initiatives such as the Digital Education Action Plan, European Media Literacy Week, and the AVMSD, the challenges posed by the rise of AI-generated media are considerable. Addressing these challenges will require a coordinated effort involving policymakers, educators, technology companies, and civil society to ensure that media literacy initiatives remain effective and relevant. As the digital landscape continues to evolve, so must the strategies and tools used to educate and protect European citizens, making media literacy a dynamic and ongoing priority.

2 National mitigation strategies

Within Europe, regulation may be a matter of shared responsibility between European institutions, national institutions and/or individual actions, depending on the issue at hand. The following section aims to provide an overview of the measures taken by European countries, both Member States and relevant regional actors, to tackle AI-generated disinformation, particularly deepfakes. The measures surveyed include national policies on generative AI, legislation aimed at harmful online content, and cybersecurity strategies aimed at mitigating the spread and/or impact of disinformation and coordinated influence operations. This review does not cover all European states but is instead limited to several regions (Western Europe, the Baltic States, and Western Balkans) in order to provide a broader overview of different approaches and issues that arise across the continent. Table 1 provides an overview of the major actions taken by each state to combat disinformation.

Table 1. Summary of state actions to mitigate disinformation

State	Actions
France	Law Against Manipulation of Information Bill on the Fight Against the Manipulation of Information Avia Law
Germany	Network Enforcement Act Interstate Media Treaty
Italy	Articles from the Italian Criminal Code Law no. 70 on Prevention and Fight Against Bullying and Cyberbullying Law no. 90 on Cybersecurity
Spain	Organic Law on the Protection of Personal Data and Guarantee of Digital Rights National Security Strategy Protocol to Counter Disinformation
United Kingdom	Online Safety Act National Security Strategy Online Media Literacy Strategy RESIST 2 Counter Disinformation Toolkit
Albania	Law no. 25 on Cybersecurity National Cyber Security Strategy
Bulgaria	Personal Data Protection Act National Cybersecurity Strategy
Greece	Coordinating Committee for AI Committee of Data Protection Officers National Transparency Authority
North Macedonia	Macedonian Fund for Innovation and Technological Development Plan for Resolute Action Against the Spread of Disinformation Strategy for Building Resilience and Tackling Hybrid Threats
Serbia	Law on Public Information and Media Law on Public Service Broadcasting Law on Electronic Media Law on Information Security Information Society and Information Security Development Strategy
Estonia	e-governance services
Latvia	Latvian AI Strategy National Research Programme
Lithuania	National Research Centre for AI

2.1 Western Europe

2.1.1 France

France's evolving legal framework on disinformation reflects a complex balancing act between preserving democratic integrity and upholding fundamental rights. The country's efforts reflect a broader trend within the European Union, where Member States are grappling with similar issues as they shape new regulations such as the DSA. However, France's approach is distinct in its willingness to experiment with robust regulatory frameworks and could serve as both a cautionary tale and a blueprint for other nations seeking to address the challenge of disinformation without impacting personal freedoms.

Law Against the Manipulation of Information

The cornerstone of France's legislative efforts against disinformation is Law Against the Manipulation of Information enacted in 2018 which primarily targets the spread of false information during election periods (Journal officiel de la République française, 2018). Colloquially known as the "fake news law", the legislation empowers judges to order the removal of false content, particularly when it is disseminated through social media platforms. While the intent behind this law is to safeguard democratic processes, its implementation has raised concerns about potential infringements on freedom of expression. Critics argue that the short timeframes for judicial decisions may lead to hasty judgments, potentially resulting in the suppression of legitimate speech (Belli, 2020). However, proponents contend that the law's focus on demonstrably false information provides a necessary bulwark against electoral manipulation (Dunn, 2024). It has been further argued that such measures are necessary to protect democracy from electoral manipulation while balancing press freedom (Ricci 2018; Young 2018). The law also imposes obligations on digital platforms to ensure transparency regarding the sources of information and to cooperate in combating disinformation (Blocman 2019). The law's relevance to AI-generated media, such as deepfakes, is implicit rather than explicit. The broad definition of false information within the legislation could encompass AI-generated content designed to mislead voters. However, the lack of specific provisions addressing the unique challenges posed by deepfakes may limit its effectiveness in this continuously evolving domain (Villasenor, 2019).

Bill on the Fight Against the Manipulation of Information

Further extending its regulatory scope, the proposed Bill on the Fight Against the Manipulation of Information granted new powers to the High Audiovisual Council (CSA), allowing it to prevent, suspend, or terminate the broadcasting of television services controlled by foreign states if they are found to infringe on France's fundamental interests (Ajji, 2020). Introduced in April 2021 as part of a broader effort to enhance France's regulatory framework in response to challenges posed by digital content including piracy and disinformation, this bill ultimately led to the formation of the Regulatory Authority for Audiovisual and Digital Communication (ARCOM) (Osborne Clarke 2021). This new authority is empowered to take significant action against audiovisual content that threatens national interests, particularly from foreign-controlled broadcasters, thereby strengthening France's capacity to safeguard its media landscape and public interests. This move

signifies a heightened awareness of the threats posed by foreign state-controlled media in the spread of disinformation and underscores the need for robust oversight mechanisms to protect national security and democratic integrity. Additionally, the bill imposes an obligation on internet service providers and hosts to enable users to flag content they believe to be fake, thus involving the public in the fight against disinformation. There is also a mandate for transparency regarding the relationships between online platforms and advertisers, aiming to curb hidden influences and ensure accountability in digital advertising practices.

Avia Law

Named after its sponsor Laetitia Avia and enacted in 2020, the Avia Law aimed to address the spread of online hate speech and disinformation by requiring digital platforms to remove what it defined as “manifestly illegal” content within 24 hours of notification. However, it faced significant backlash for potentially encouraging over-censorship and infringing on the right to freedom of expression (Marotta, 2020). The French Constitutional Council declared key provisions of the law unconstitutional on June 18, 2020, criticising the lack of judicial oversight and the risk of excessive content removal driven by tight deadlines and hefty penalties. While the ruling from the French Constitutional Council invalidated several of the more controversial provisions of the Avia Law, particularly those mandating the rapid removal of "manifestly illegal" content without judicial review, it also allowed for the continued enforcement of certain remaining provisions. These provisions can still be applied to specific types of disinformation that cross legal boundaries, such as incitement to violence or hate speech. This decision has broader implications for France’s strategy to regulate digital content and influences ongoing debates within the European Union. It highlights the difficulty of crafting laws that effectively combat harmful online content without undermining freedom of expression. The ruling points to the need for a balanced approach that respects fundamental rights while addressing the challenges posed by new forms of disinformation, including AI-generated content. As France’s model for digital content regulation faces scrutiny, it serves as a crucial case study for European lawmakers who are navigating similar issues. The Constitutional Council’s decision underscores the necessity of a nuanced regulatory framework that can address evolving threats while upholding core democratic values.

2.1.2 Germany

While Germany largely follows EU legislation, there is no specific legislation concerning the production and distribution of AI-generated content (e.g., deepfakes) online. As there is no explicit requirement to identify or label digitally manipulated content, there is a vulnerability with respect to AI-generated disinformation. While there is currently a discussion ongoing around the need for a more detailed legal framework, existing legislation relies on laws already in place that protect privacy, dignity and security.

Network Enforcement Act

Enacted in 2017, the Network Enforcement Act (NetzDG) introduces strict obligations for internet companies and the handling of harmful content that is hosted on online platforms. Notably, NetzDG requires online

platforms to introduce an effective and transparent procedure for handling removal complaints that concern illegal content and an obligation for such platforms to publish a transparency report on a biannual basis. Furthermore, the Act requires online platforms with more than 2 million registered users in Germany to exercise a local takedown of "obviously illegal" content within 24 hours after notification. To qualify for removal under NetzDG, content must fall under one of the 21 criminal statutes in the German Criminal Code (StGB). Where the legal status of such content is unclear, the provider normally has up to seven days to decide on the case. This can take longer in exceptional circumstances, for example, if users that uploaded the content are asked to weigh in or if the decision is passed to a joint industry body accredited as an institution of co-regulation.

Interstate Media Treaty

Enacted in November 2020, the Interstate Media Treaty (MStV) replaces the previous State Broadcasting Treaty and transposes the revised AVMSD into national legislation but also addresses other elements of the German media landscape. In particular, it contains provisions regarding media platforms, user interfaces and media intermediaries. In relation to the implementation of the AVSMD, the MStV contains provisions to protect minors and human dignity on video-sharing platforms, to strengthen barrier-free servies and to relax advertising restrictions for private broadcasters. As regards to so-called "new media providers", it also establishes general principles in the form of technology-neutral rules, transparency obligations and non-discrimination requirements. It contains specific regulations for media platforms (provisions on signal integrity, accessibility and discoverability rules) and media intermediaries (obligations to label "social bots" and to nominate an authorised agent in the host country).

2.1.3 Italy

Italy is among many states that are attempting to tackle the issue of AI-generated disinformation specifically in the contexts of identity theft, freedom of information, cybersecurity and pornography. Although Italian laws, such as those detailed below, can generally be applied to a large variety of cases, there remains a lack of legislation that clearly and specifically addresses the production and dissemination of AI-generated disinformation.

Articles of the Italian Criminal Code

The Italian Criminal Code, also known as the Rocco Code, came into force in the 1930s and has since undergone significant changes to keep up to date with technological and cultural developments. Notably, Article 656 of the Code criminalizes the publication or dissemination of "false, exaggerated or tendentious" news that is likely to disturb public order. Within this definition, news that is completely dissimilar to reality must be considered false and so Article 656 may extend to particularly harmful AI-generated disinformation. However, the definition of criminalized news provided in Article 656 may be broadly interpreted and, as the protected legal asset is public order itself, the level of impact that warrants criminalization is also ambiguous.

Article 595 on defamation is also of relevance to the production of AI-generated content, particularly deepfakes.

In 2019, the Criminal Code was further amended to address the production and distribution of revenge porn or non-consensual pornography. Article 612 criminalized the act of publishing sexually explicit images or videos, intended to remain private, without the consent of the persons involved and this offence is considered committed instantaneously upon first sending such content. The Criminal Code defines the offence as the "unlawful dissemination of sexually explicit content, which may have as its object images or videos depicting sexual acts or genital organs or even other erogenous parts of the human body, such as breasts or buttocks, naked or in conditions and context such as to evoke sexuality". Even though AI-generated pornographic images are artificial, this amendment indirectly applies to their dissemination. However, it has been proposed that Article 612 should be amended further to explicitly apply to AI-generated images.

Law no. 70 on Prevention and Fight Against Bullying and Cyberbullying

Enacted in May 2024, Law No. 70 contains provisions aimed at preventing and countering cyberbullying. Notably, these provisions aim to increase resources available for prevention and awareness-raising information campaigns, promote initiatives that provide psychological support services for students in schools, provide internal codes and obligations for school staff to apply appropriate procedures to tackle cyberbullying, and to promote broader educational initiatives. The production and dissemination of AI-generated content featuring an individual's likeness, particularly in sexual contexts, constitutes a clear form of cyberbullying that this law seeks to address.

Law no. 90 on Cybersecurity

Enacted in June 2024, Law No. 90 aims to strengthen Italy's cyber resilience by tightening penalties for cybercrimes on the one hand and strengthening prevention and law enforcement tools on the other. Among other provisions, the legislation introduces significant increases in penalties for offences such as unauthorised access to a computer system or the damaging of computer information, data and programs. This new legislation provides for the introduction of common aggravating circumstances and some special aggravating circumstances in cases where offences are committed with the use of AI systems but also envisages the introduction of a new offence related specifically to deepfakes.

2.1.4 Spain

The Spanish disinformation landscape is characterized by intense political and media polarization, often manifesting through narratives that blend real data with distortions or outright falsehoods. Malicious actors exploit genuine public concerns by layering multiple narratives within a single hoax and recycling them across different crises, making it increasingly difficult for the public to distinguish fact from fiction. Recent years have seen Spain targeted by various disinformation campaigns aimed at destabilizing democratic processes, frequently originating from unknown or foreign sources but later amplified by domestic actors. Examples of

such disinformation incidents in Spain include fabricated claims linking the Catalan independence movement to Russian financial support, which served to delegitimize the movement while fostering political strife within the country. Additionally, during the migrant crisis in the Canary Islands, misleading narratives circulated about Spain financially supporting "illegal immigration" and an alleged invasion of young male migrants, further fuelling anti-immigrant sentiment and xenophobia (Romero Vicente, 2023). The COVID-19 pandemic also saw the rapid dissemination of false information, including outright denials of the virus's existence, promotion of ineffective cures, and the spread of conspiracy theories that undermined public trust in health measures. These narratives misled the public and complicated efforts to manage the health crisis .

Spain's approach to addressing misinformation highlights both strengths and limitations, particularly in the context of widespread disinformation. The National Security Strategy recognizes disinformation as a significant security threat, elevating it within government priorities and facilitating coordinated responses among various agencies during information crises. This collaborative framework involves national security, communication, and media oversight bodies working together to counteract misleading information and protect public discourse. However, deep political and media polarization complicates these efforts, as competing narratives can be amplified by both traditional and social media, making it easier for malicious actors to spread disinformation. Spain's strategy seeks to balance reactive measures, such as immediate responses to misinformation incidents, with proactive initiatives focused on enhancing public resilience through media literacy and critical thinking education. Given the rapid evolution of disinformation tactics, the government acknowledges the necessity for adaptable strategies to effectively address new threats in an increasingly digital landscape. With the rise of new technologies, including deepfakes and AI-generated content, the government faces the task of staying ahead of these threats while ensuring the protection of democratic integrity. Therefore, a multifaceted approach that combines regulatory measures, public education, and inter-agency cooperation is essential for effectively addressing the challenges posed by disinformation in Spain.

Organic Law on the Protection of Personal Data and Guarantee of Digital Rights

Spain's primary legislative effort in the digital sphere is the Organic Law 3/2018 on the Protection of Personal Data and Guarantee of Digital Rights. While this law primarily focuses on data protection, it also includes provisions that indirectly address issues related to disinformation (Boletín Oficial del Estado, 2018). Article 85 of this law establishes the right to rectification on the Internet, allowing individuals to request the removal or correction of inaccurate information about them online. This provision, while not explicitly targeting disinformation campaigns, provides a mechanism for combating personal disinformation (Recio Gayo, 2019). While the law's relevance to AI-generated media is limited, the right to rectification could potentially be invoked in cases where AI-generated content contributes to false or misleading information about individuals.

National Security Strategy

Updated in 2021, Spain's National Security Strategy identifies disinformation as a significant threat to national security and emphasizes the need for a coordinated response to disinformation campaigns, particularly those

originating from foreign actors, threatening the internal stability of the country (Goberno de España, 2021). The strategy informs policy development and implementation across various government agencies and called for enhanced cooperation between public institutions, media organizations, and technology companies to combat the spread of false information. While the strategy acknowledges the role of emerging technologies such as AI systems in the production and dissemination of disinformation, it does not provide specific regulatory measures for addressing the spread of AI-generated content.

Protocol to Combat Disinformation

In response to concerns about electoral interference, the Spanish government introduced a Protocol to Combat Disinformation in 2019. This initiative establishes a coordination mechanism between various government agencies to monitor and respond to disinformation campaigns, particularly during election periods (Boletín Oficial del Estado, 2019). The protocol emphasizes rapid response and inter-agency cooperation, but it has been criticized for its potential to infringe on freedom of expression. Critics argue that the broad definition of disinformation used in the protocol could lead to the suppression of legitimate political discourse. While the protocol does not explicitly address AI-generated media, its focus on rapid identification and response to disinformation may be applied to deepfakes and other AI-generated content that poses a threat to electoral integrity.

In Spain, the legal consequences of spreading disinformation are broad and vary significantly depending on the content and intention behind its dissemination. The spreading of disinformation can lead to criminal charges ranging from hate crimes and offences against moral integrity to threats to public order, defamation, or slander. It can also encompass crimes against public health, fraud, and unauthorized practices particularly in the context of unproven and ineffective alternatives cures which represent a growing segment of the disinformation landscape. In 2022, Spain saw its first conviction for spreading disinformation that incited hatred, marking a significant legal precedent. The case involved a user who posted a video of an assault, falsely accusing a foreign minor as the perpetrator. The user received a sentence of 15 months in prison and a fine of 1,620 euros, reigniting discussions about criminal liability for those who disseminate false information.

Initiatives to counter disinformation

Spain actively encourages technology companies operating within its borders to adhere to the EU's Code of Practice on Disinformation and to take significant action against the spread of false information, recognizing the need for a unified effort in combating disinformation. By promoting the principles outlined in the code, Spain seeks to enhance the integrity of public discourse and protect democratic processes in an increasingly digital landscape. Spain's response to misinformation is also shaped by a blend of independent fact-checkers, media-integrated efforts, and government initiatives. Independent organizations such as Maldita.es, Newtral, Verificat, and Infoveritas play a crucial role in debunking false information. Media outlets like EFE Verifica and AFP Factual also contribute to fact-checking efforts. Government initiatives, including the forum against disinformation campaigns and the PSOE's plans to address misinformation in the context of upcoming elections, demonstrate a comprehensive approach to tackling false narratives (PSOE, 2023).

2.1.5 United Kingdom

In contrast to other countries in Europe highlighted in this section, the UK is no longer a Member State of the EU and does not adhere to EU legislation such as the AI Act and DSA. However, the UK has introduced legislation to fulfil similar goals of mitigating the spread and impact of online disinformation, notably the Online Safety Act, as well as cybersecurity and media literacy initiatives for tackling disinformation.

Online Safety Act

Introduced in 2023, the Online Safety Act assigns online platforms with a duty of care to their users such that these companies are required to promote online safety by taking action against illegal and harmful content circulating online. Primarily, this duty of care is designed to ensure the protection of children online but the Act also requires platforms to provide adult users with optional tools for reducing the likelihood of viewing unwanted content including offensive material. Notably, the Act introduces provisions that specifically combat the spread of disinformation and online harassment (UK Government 2023). Among several new criminal offences introduced, the Act includes a false communications offence for users that knowingly send false content online with the intention of causing non-trivial psychological or physical harm. This offence extends to harm arising from the nature of the content, the fact of its dissemination, and the manner of its dissemination. In this context, harm may refer to instances where the sending of such content results in individuals causing harm or increasing the likelihood of causing harm to themselves or others. The Act further criminalises the sending of deepfake pornography by deeming it an offence to intentionally send a photograph or film of any person's genitals to an unwilling recipient, including an image that appears to be a photograph or film by that has been made or altered by computer graphics or in any other way. Additionally, the Act establishes a committee to advise the UK media regulator, the Office of Communications (Ofcom), on the implementation of its Media Literacy Strategy.

In spite of these provisions, the Online Safety Act has been criticized for not going far enough to effectively tackle the harms of disinformation and threats of AI-generated media in particular. Notably, unlike the DSA, the Online Safety Act defines harm narrowly as physical or psychological harms resulting from individual actions, rather than considering broader societal harms resulting from the online ecosystem itself (Farrand 2024). This narrow scope means that Ofcom has limited oversight and enforcement emphasizes content takedown rather than tackling broader issues such as platform architecture that incentivizes or facilitates the spread of harmful content (Judson et al. 2024). Additionally, the criminal offences introduced in the Online Safety Act have been criticized as being insufficient to counter non-consensual pornography, particularly deepfake pornography (Kira 2024). Notably, the Act does not directly address election disinformation unless instigated by foreign actors and is designed to address the publication and spread of individual content rather than disinformation at scale. As such, the Act would likely be ineffective at combatting coordinated disinformation campaigns, conspiracy theories and information incidents in which a significant amount of false content spreads rapidly and widely online (Full Fact 2024). Finally, the regulatory

differences between the DSA and the Online Safety Act may limit collaboration and the development of cross-border strategies that are necessary to tackle widespread disinformation (Farrand 2024).

Cybersecurity initiatives

In 2021, the UK government launched its National Cyber Strategy 2022 setting out plans for protecting and promoting the country's interests in cyberspace (UK Government 2022). While the National Cyber Strategy does not explicitly discuss AI-driven disinformation, these plans do include strategies for defending democratic institutions and processes from foreign manipulation through disinformation campaigns. In order to combat foreign interference, including disinformation campaigns, the National Cyber Strategy 2022 proposes a whole-of-society approach that recognises cybersecurity as something that needs to be integrated into all forms of policymaking and that cannot be addressed by central government alone. While government is responsible for setting and enforcing laws and standards and providing technical guidance, large technology companies are also equally responsible for maintaining a secure cyber environment and businesses and organizations are responsible for managing cyber risks and protecting data. Primarily, the National Cyber Strategy proposes a decentralized approach to cybersecurity with a UK-wide network of regional cyber clusters and cyber resilience centres, better linking local organizations, academic centres and businesses with local law enforcement, as well as with Government Communications Headquarters (GCHQ), the UK's national intelligence agency. Broadly, the whole-of-society approach taken by the National Cyber Strategy 2022 has been received positively (Clark 2024).

In regards to the role of AI in cybersecurity, in 2021 the UK government launched its National AI Strategy outlining long-term plans for developing the country's AI ecosystem with a focus on encouraging innovation and investment, while also seeking to protect the fundamental rights of UK citizens against potential harms (UK Government 2021). The National AI Strategy specifically notes the need to address AI risks including the use of deepfakes and AI-driven disinformation, particularly from foreign actors. While the National AI Strategy does not introduce specific measures or regulations to combat AI-driven disinformation, in 2023 the government launched the state-backed AI Safety Institute with the goal of researching AI safety measures including societal implications such as infodemics. In comparison to the EU AI Act, the UK's National AI Strategy has been criticized for its focus on utilizing the benefits of AI in future economic and technological development rather than pursuing ethical AI development and deployment or introducing strict regulations and policies for AI governance (Montasari 2023).

Online Media Literacy Strategy

In 2021, the UK government published its Online Media Literacy Strategy established by the Online Safety Act and to be implemented by the media regulator Ofcom (DCMS 2021). The Strategy sets out a coordinated approach to improving the UK public's media literacy skills by supporting citizens and organisations to undertake media literacy activities through various initiatives. At its core, the Strategy outlines a Media Literacy Knowledge and Skills Framework that internet companies can follow in order to promote the media literacy of their users. This Framework highlights that companies should take steps to ensure that users

understand how their personal data can be exploited online, how they can protect their privacy, how to critically analyse online content, how actions online have real-world impacts, and how to participate positively in online environments. While the Online Safety Act places specific legal requirements on internet companies, the principles set out in the Online Media Literacy Strategy are entirely voluntary. The Strategy further proposes that online platforms implement "literacy by design" measures to address the spread of disinformation (e.g., fact-checking content, nudging prompts).

Furthermore, the UK government has also published a Media Literacy Resources Hub online with links to third-part sites for media literacy training and resources.

The Online Media Literacy Strategy does highlight information literacy as a subset of media literacy and focuses efforts on critically engaging with mis/disinformation online but does not highlight the risks of AI-generated media. Neither do either of the annual action plan updates highlight increasing risk of AI-generated media.

RESIST 2 Counter Disinformation Toolkit

Updating the original RESIST framework introduced in 2018, the RESIST 2 framework provides organisations with an extensive step-by-step strategy or checklist for responding to disinformation attacks (GCS 2022a). This framework is broken down into multiple steps:

- *Recognise*: outlines how to determine if a piece of information is false, how best to distinguish between misinformation and disinformation and why these categorisations are useful for understanding the different impacts on audiences.
- *Early warning*: outlines how organizations can identify areas vulnerable to mis/disinformation and how best to monitor these vulnerabilities and the extent to which they are at risk.
- *Situational insight*: outlines how organizations may gain insights into and respond to emerging threats and issues, as well as how best to communicate these insights clearly to officials.
- *Impact analysis*: outlines how organizations can use structural analysis techniques to predict the likely impact of a piece of mis/disinformation, emphasizing the need for clearly defined processes for assessment.
- *Strategic communications*: outlines how organizations can develop and disseminate content that is engaging, impactful and uses "friendly voices" to increase credibility and reach, as well as how organizations can develop different responses in diverse scenarios.
- *Tracking effectiveness*: outlines how best to measure the effectiveness of strategic communications and offers examples of metrics that can be used to assess if communications have achieved pre-defined objectives.

RESIST 2 does not introduce specific measures for combatting AI-generated media but it assumed that this framework accounts for and can respond to AI-driven disinformation.

The Wall of Beliefs

The Wall of Beliefs resource for civil servants is a toolkit that advises on understanding false beliefs and provides 4 different strategies for effectively countering disinformation (GCS 2022b). The titular "wall of beliefs" is a model for understanding how beliefs are linked and interdependent, how they are established through our social environment, and how difficult different kinds of beliefs are to change. According to this model, different kinds of belief are more deeply embedded than others. From least to most deeply embedded, beliefs can be arranged as follows: beliefs derived from fashions, beliefs derived from recent news, beliefs derived from norms, beliefs derived from trusted sources, foundational beliefs, and sacred beliefs. Depending on how embedded a false belief is and the extent to which it may cause harmful behaviours, the toolkit proposes 4 different strategies for countering these beliefs: proactive promotion, manage behaviours, reactive response, and watch and wait. The toolkit further provides a step-by-step procedure for how to develop an appropriate response plan for (i) when false beliefs or disinformation narratives are detected and (ii) when harmful belief-driven behaviours are detected. As the toolkit acknowledges, this framework is limited and does not fully account for the role of technological infrastructure, platforms and tools such as emerging generative AI systems.

2.2 Western Balkans

2.2.1 Albania

Albania is one of the first countries in Europe to transpose Law No. 25/2024 "On Cyber Security" as a commitment towards membership in the European Union. This comprehensive piece of legislation was introduced to ensure that the country is prepared to respond to ongoing cybersecurity threats and developing situations. The law includes provisions for preparing critical information infrastructures to withstand cyberattacks and further provides procedures for reporting and handling cybersecurity incidents in an appropriate and coordinated manner. In 2020, the Albanian government also approved a new five-year National Cybersecurity Strategy 2020-2025 that was crafted within the framework of the National Security Strategy 2014-2020 and that lays out the nation's strategy for enhancing its cybersecurity landscape. It encompasses key principles like safeguarding fundamental rights, upholding freedom of expression, protecting personal information and privacy, ensuring access for all, promoting democratic and effective governance, and fostering collective responsibility in upholding cybersecurity. Additionally, the strategy places special emphasis on the protection of children in the online sphere.

There is no specific regulation on disinformation in Albania, nor direct use of this term or a legal definition of what constitutes disinformation. However, the Criminal Code regulates the dissemination of false information with the aim of causing panic in Article 267, which states: "Spreading false information or news, in words, in writing, or in any other manner, in order to incite a state of insecurity or panic in people, is punishable by a fine or up to five years of imprisonment." Furthermore, Article 271 covers cases in which disinformation is given to emergency units intentionally to hinder their efficacy, committed by any means of information or communication, and is a contravention punishable by a fine or up to one year of imprisonment.

Despite some institutional preparedness in terms of legislation, academics around the world note that Albania ranks towards the lower quadrants in the Western Balkans region when it comes to readiness for implementing advanced technologies such as generative AI. Notably, Albania has been ranked 85th among 172 countries in the Government AI Readiness Index based on various indicators including digital capacity, data availability, and data representativeness and 80th among 180 countries in the Network Readiness Index (Krivokapić, Živković and Nikolić 2022). These readiness measurements suggest that Albania may be less prepared for the arrival of generative AI and so it is more likely that AI-generated disinformation will have a disruptive effect on societal stability.

Another area of concern is Albania's media landscape which has been repeatedly characterized by local researchers as "weak, fragmented and constantly changing" due to political and financial pressures, as well as transparency issues, ethical concerns and quality issues (Kokalari 2024; Bieber & Kmezic 2015; Kapo 2023). However, it is difficult to monitor Albania's media landscape because it is rapidly changing and overcrowded with over 30 media outlets in the country, thus making the country especially vulnerable to the spread of disinformation. This is evidenced by, for example, Russian disinformation campaigns that have occurred in the Western Balkans and the inflammatory role media played in vaccination debates during the Covid-19 pandemic (Nenovski, Ilijevski, Stanojoska 2023). Therefore, while there is sufficient regulation on media freedom, rights of journalists, protection of personal data, and defamation laws in Albania, there remains a challenge in enforcing these laws and ensuring transparency for media organizations.

Ultimately, Albania lacks effective disinformation mitigation policies and this issue is only exacerbated by a lack of digital literacy in the country, particularly around generative AI. While there have been efforts to tackle this issue through workshops and public discussions on AI, these have not translated into formal regulatory measures. Notably, Prime Minister Edi Rama has made propositions to integrate AI technologies into public services like e-Albania but the current lack of a cohesive strategy raises concerns about the potential misuse of AI and the public's understanding of these technologies (Tufa 2023). Some researchers have criticized current policies as being superficial in that they only provide some tangible but short-term benefits (Kokalari 2024). These policies may be enough to satisfy the electorate for a time but they are hindered by a weak infrastructure and a dysfunctional system (Ibid.). Finally, the government is the only important actor in Albania currently working to mitigate AI-generated disinformation but further organizations could be mobilized to tackle this issue such as the Albanian Institute for Safe AI, Albania Artificial Intelligence, and the National Agency for Information Society, as well as relevant technology companies.

2.2.2 Bulgaria

While Bulgaria has put in place some regulatory provisions to tackle both AI and disinformation issues, these remain fragmented. In seeking to uphold and supplement current EU data protection laws, Bulgaria has enacted the Personal Data Protection Act which addresses the issue of data abuse in the application of any technologies related to AI. Furthermore, in 2020 the Bulgarian government adopted the Digital Bulgaria 2024 plan that aimed to enhanced and apply the use of information technologies in all areas of society. This strategy stresses

the need for ethical principles regarding the use of AI systems in an effort to prevent violations such as data misuse, lack of accountability and lack of transparency in algorithms. Additionally, the Ministry of Electronic Governance has also implemented several programs with the objective of enhancing the digital capabilities and uses of AI technologies within public administration.

Adopted in 2019, Bulgaria's National Cybersecurity Strategy aims to protect critical information infrastructure, to make this infrastructure more resilient against cyber threats, and to improve the cybersecurity provisions of public and private entities more broadly. The strategy emphasizes the need for public awareness and education for cybersecurity, the need to invest in technology and research, the importance of domestic and international cooperation on these issues, and the implementation of a National Cybersecurity Incident Response Team.

While some progress has been made to ensure Bulgaria is resilient against AI-driven disinformation, the media landscape in Bulgaria also faces significant challenges relating to credibility, trust and political and economic influences (Vasileva 2023; Margova & Dobreva 2023). Acknowledging this, the Bulgarian government has initiated an Action Plan for Combating Disinformation that aims to enhanced media literacy, support journalistic independence and improve mechanisms of fact-checking.

Ultimately, more proactive efforts to specifically tackle the role of AI in disinformation and the ethical use of AI are required from regulatory bodies in Bulgaria.

2.2.3 Greece

Greece has taken significant steps toward addressing disinformation and the possible political impacts of AI-generated content. Notably, in March 2023 the Greek government introduced new legislation on emerging information and communication technologies, strengthening digital governance and other provisions. This new legislation outlines its aim to provide appropriate guarantees to ensure the rights of natural persons and legal entities, strengthening accountability and transparency in the use of AI systems, and complementing the existing institutional framework for cybersecurity. The legislation further outlines obligations for public bodies that use AI systems to be transparent and publish publicly available information on the system start-up time, the operating parameters and capabilities of the AI system, as well as the categories of decisions taken by or with the support of the AI system. The legislation also provides mandatory conditions for the provision of AI services to public bodies including that the public body may study the way the AI system operates, taking into account the intended purpose of the system, among other things. Likewise, under this new legislation every public sector body is required to keep a record of AI systems used that is updated annually.

This new legislation also established the creation of a Coordinating Committee for AI responsible for the application of Greece's National AI Strategy and also an AI Observatory in the Ministry of Digital Governance responsible for data collection on the National AI Strategy, drawing up key indicators on AI activities in Greece, best practices for the use of AI in the private and public sectors, and the effects of AI on the fundamental rights of natural persons. It further introduced a rigorous structure for the use of AI systems in public bodies with provisions to safeguard the protection of the rights of natural or legal persons affected

by these systems. This includes a Committee of Data Protection Officers and a National Transparency Authority.

The Greek government is taking significant steps in establishing AI policies but there have been calls from academics and experts for a more integrated approach that focuses on public education, transparency and accountability. Academics note that the risks associated with disinformation grow as AI technology evolves, particularly in political and social contexts. They stress the need to conduct rigorous research and the need for ethical consideration of how AI systems can serve the public good and inform public discourse, and further advocate for the implementation of media literacy programmes that give citizens the tools needed to critically evaluate information online (UNESCO 2024; Infinite 2023). Policy recommendations from thinktanks prioritise multi-stakeholder engagement and further highlight the importance of promoting responsible AI usage to prevent disinformation rather than simply penalizing AI misuse (Tsamados 2021). Similarly, experts have supported the use of AI tools for identifying and mitigating false information in real time, while also considering the privacy violation and algorithmic bias that arises from overreliance on AI.

2.2.4 North Macedonia

North Macedonia currently lacks AI-specific regulations but there is a strong commitment to developing a comprehensive AI strategy that will align with EU standards. The most significant initiatives for developing and implementing AI and other advanced technologies in North Macedonia are closely tied to the government and the Macedonian Fund for Innovation and Technological Development (FITD). However, the Ministry of Information Society and Administration serves as the main coordinator of all digital-related activities.

The FITD and the government initiated efforts in 2021 to create a National AI Strategy, however, progress has been slow due to challenges such as insufficient data, human resources and technical capabilities. Notably, data privacy is a major concern for AI development given the technology's ability to process vast amounts of data. While the Macedonian Data Protection Act aligns with the EU GDPR, it may not sufficiently address privacy challenges posed by AI. This is further complicated by the fact that intellectual property issues related to AI remain unresolved in North Macedonia and there is concern that the use of vast datasets for AI learning may infringe intellectual property rights. Future amendments and legal developments in these areas are expected to more closely align North Macedonian laws with the EU to ensure comprehensive privacy protection.

Disinformation poses a considerable threat to North Macedonia that may impact political and social dynamics, as well as public health and safety. The current government has recognized this threat and has made the fight against disinformation a priority since 2019 when the government launched the Plan for Resolute Action Against the Spread of Disinformation. While this plan does not explicitly mention the role of generative AI in disinformation, it does aim to address the challenges of disinformation more generally and advocates for multi-stakeholder engagement (government, civil society, media organizations and international partners) to tackle the issue. Notably, the plan proposes the need for media literacy initiatives, public awareness campaigns,

monitoring and reporting mechanisms for disinformation campaigns, as well as proposing the development of legislative frameworks that are currently non-binding but actively promote responsible media practices.

In October 2021, the North Macedonian government made additional efforts to address disinformation and hybrid threats more broadly by adopting a Strategy for Building Resilience and Tackling Hybrid Threats. This was then followed by a 2021-2025 action plan that makes use of preliminary oversight activities and suggests the utilization of communication channels between informal parliamentary groups and civil society.

Despite positive efforts being made to tackle disinformation, many experts continue to criticize the lack of enforceable measures that may see the plans fall short in effectively addressing disinformation challenges. Academics on the other hand stress the need for stronger literacy in order to strengthen citizen participation in recognizing and fighting disinformation (Kokalari 2024; Bieber & Kmezic 2015; Kapo 2023). Moreover, there have been independent analyses that call for policies which target online platforms specifically, which have been characterized as catalysts for spreading disinformation. Also, while AI is not specifically mentioned in national plans and legislations, academics and think tanks are calling for inclusion of these new technologies in the national plans, especially regarding literacy and ethical considerations surrounding its use in media. Finally, some experts maintain that a regional approach to disinformation that includes collaboration and shared resources among neighbouring countries may be needed, as North Macedonia's policies still lack depth (Kokalari 2024; UNDP 2023).

2.2.5 Serbia

There is no specific piece of legislation that forbids the spreading of disinformation in Serbia. However, the Law on Public Information and Media prescribes that "the rules on public information provide and protect the release, receipt and exchange of information, ideas and opinions through the media with a view to improving the values of a democratic society, preventing conflict and preserving peace, authentic, timely, reliable, and complete informing and enabling free personal development." (Article 2, Law on Public Information and Media). In addition, the Law states that media have "to get true, complete and timely information about the issues of public importance and the means of public importance shall honor that right" (Article 5, Para 2 of the Law on Public Information and Media). Furthermore, Article 9 prescribes that the editor-in-chief of the publication and the journalist have the obligation "prior to publishing the information about an occurrence, an event or a person, to check its origin, authenticity and completeness with due diligence appropriate for the circumstances." Additionally, the Law on Public Service Broadcasting states that public service media organizations have a special obligation to obey the public interest through "authentic, timely and full information available to all citizens of the Republic of Serbia" (Article 17, Law on Public Service Broadcasting), while the Law on Electronic Media obliges providers of media services to enable "free, accurate, objective and timely information" (Article 47, Law on Electronic Media).

Serbia's legislative framework on cybersecurity is structured upon the existing Law on Information Security that was introduced in 2016 and which primarily focuses on measures to safeguard information and communication systems against security risks (RS Official Gazette 6/2016, 94/2017 and 77/2019). The Law

further delineates the responsibilities of legal entities in managing and using these systems, as well as the competent authorities tasked with implementing protection measures. The Law ushered in several institutional developments, including the creation of a national Computer Emergency Response Team (CERT) within the Serbian Telecommunications Agency (RATEL) that "monitors the status of incidents on national level, provides early warnings, alerts and announcements, and reacts on incidents by providing the information on affected entities and persons" (REF). The CERT also holds seminars and technical trainings for operators of special importance (critical infrastructure). There have been several key amendments to the Law. Firstly, the purview of information security has been extended to cover not only the state but also the economy and local self-government underscoring the comprehensive nature of this legislation. Secondly, the Law has been amended to create an Office for Information Security which will oversee various levels of information security and manage incidents effectively in what is a critical step toward ensuring a coordinated response to security threats. A final amendment has also tasked the CERT to maintain a Vulnerability Database for ICT products and services that will allow both natural and legal persons, as well as manufacturers, supplies, and service providers within the ICT system to voluntarily report vulnerabilities. This initiative will facilitate the identification and mitigation of security weaknesses in the information security landscape. The Information Society and Information Security Development Strategy 2021-2026 provides a comprehensive approach to information security for ICT systems of special importance to the security of Serbia, as well as for private citizens and businesses.

In light of EU accession negotiations, Serbia has signed and ratified the Council of Europe Convention on Cybercrime (Budapest Convention), including its Additional Protocol on Xenophobia and Racism Committed Through Computer Systems. Nevertheless, the country also voted in favour of UNGA resolution 74/247 calling for an international legal instrument to govern this domain. Domestically, the national legislative framework has been developed in accordance with the Budapest Convention and EU legislation. A High-Tech Crime Unit has recently been established within the special prosecutor's office, along with three specialized units: crime analysis; terrorism and extremism; and drug prevention, addiction and repression.

Serbia stands out in the Western Balkans region for having a National Strategy for the Development of Artificial Intelligence 2020-2025. However, this strategy has encountered challenges in implementation, especially regarding transparency and accountability around government use of AI systems. Citizen trust towards the Serbian government was further challenged by data breaches during the Covid-10 pandemic which revealed significant flaws in data protection measures (Nenovski, Ilijevski and Stanojoska 2023). Serbian thinktanks have also expressed concern around implementation, noting that existing AI policies are very ambitious and thus need robust mechanisms to ensure ethical governance, protect civil rights and ensure responsible deployment of AI systems (Bjeloš & Pavlović 2022). These institutions further highlight the need to involve other actors, particularly from civil society and academia.

Ultimately, the Serbian government has aimed to enhance technological capabilities that relate to AI and disinformation through the work of various institutions including the State Center for Data Management and Storage, the Center for the Promotion of Science, Vojvodina ICT Cluster and the Institute for Artificial Intelligence of Serbia. However, Serbia has struggled with navigating civil liberty concerns while developing

these capabilities. A notable example being the Safe Society project, an initiative that planned to deploy a comprehensive video surveillance system with facial recognition capabilities in Belgrade with the Chinese technology company Huawei as a technical partner (Krivokapić, Živković and Nikolić 2022). There was no prior public debate or consideration of the social effects of its use and its announcement sparked significant civil society opposition. This difficulty in navigating civil liberty concerns and difficulty in anticipating the repercussions of using AI technologies is a common theme in policy and regulation discussions in the Western Balkans.

2.3 Baltic States

2.3.1 Estonia

Alongside the other Baltic States, Estonia has been focusing on improving its national defence in recent years in order to counter Russia's aggressive policies in the region. As such, the Estonian government has sought to build the necessary cybersecurity capacity to protect the country from digital damage and has framed this policy issue as a national effort, while further seeking to influence international policy and collaborating with other countries on these issues (Górka 2023). While a relatively small country, Estonia is internationally leading in e-government services and its cybersecurity expertise is recognized globally with other countries following their model. Beyond cybersecurity concerns, Estonia views public awareness of and technical competence in emerging technologies, particularly generative AI, as a key factor to maintaining national security. While adhering to the EU requirements such as the AI Act, Estonia's national AI strategy aims to fully harness the potential of AI by developing and implementing policy measures in areas such as encouraging the use of AI applications in both the public and private sector, providing direct support to AI research and increasing relevant skills and competencies, and developing a legal environment for AI uptake (Estonian Government 2019). To foster AI development, Estonia foresees increasing the capacity of AI research through developing AI-related research support measures and increasing the capacity and awareness of funding opportunities (Implement 2023). The uptake and development of AI in the private sector will be supported through existing funding measures such as innovation vouchers, development vouchers, and product development grants. The Estonian government is also preparing an innovation competition to promote AI developments based on governmental datasets and another funding scheme providing financial support to pilot projects at various levels of their technological readiness. The strategy also foresees reforms to formal education and training systems to increase skills and competencies in AI. Reforms at preschool, primary and secondary education will be primarily covered through the ProgeTiger Programme, while higher education reforms include Masters programmes in data science and AI, promotion of elective courses on AI in postgraduate disciplines, and increased doctoral scholarships in AI-related fields.

2.3.2 Latvia

Similarly to the other Baltic States, Latvia's strategy for addressing disinformation and AI-driven threats is greatly influenced by concerns over aggressive Russian policies and cybersecurity threats, and primarily focuses on professionalizing strategy and adapting standards to align with EU requirements (Górka 2023). With regards to generative AI, in 2020 the Latvian government released its national AI strategy which aimed to promote AI adoption and growth within the economy, identifying priority sectors with high potential for AI applications such as transport, culture, justice, agriculture and translation (Latvian Government 2020). The Latvian AI strategy includes raising awareness and competencies in AI through education reforms, promoting AI adoption in the public and private sectors, engaging in national and international cooperation, developing a solid legal and ethical framework for AI, and investing in a digital and telecommunication infrastructure. Notably, the Ministry of Interior plans to spend approximately €1.5 million on digitisation with a focus on AI and the Ministry of Culture plans to supplement machine translation systems with new language pairs to increase the availability of government content and e-services. To accelerate AI deployment, the Latvian strategy advocates for the integration of AI themes in the general education system at all levels. The strategy also plans to develop an equivalent online course for expert and management-level specialists to support digital transformation. Furthermore, the Latvian Government intends to prepare a National Research Programme and reform the education system.

2.3.3 Lithuania

Similarly to other Baltic countries, Lithuania's approach to AI and disinformation relies upon increasing the integration of this technology within the country and by improving public understanding (Górka 2023). In 2019, the Lithuanian government released its AI strategy which aimed to modernize and expand the country's current AI ecosystem while outlining several key goals including improving skills and education in AI for all citizens, strengthening national research and innovation, increasing AI deployment in all economic activities, promoting national and international collaborations in AI, developing an ethical and legal framework for sustainable AI applications, and establishing a responsible and efficient data ecosystem for AI (Create Lithuania 2019). The Lithuanian strategy presents policy recommendations for the growth of research and development in the field of AI, including establishing a national research centre in AI and increasing financial support for AI research through new funding programs. The strategy advocates for the development of skills and competencies in AI at all education levels, emphasizing the need to start teaching AI foundations at an early age. Notably, reforms to the primary and secondary education system could target AI basics for children, and more courses to develop technical skills. The Lithuanian Government recommends modernizing STEM teaching subjects and providing dedicated support to teachers to foster the quality of their education duties in AI. At the moment of publishing, there are two new undergraduate level AI programs at Lithuania's universities, and five undergraduate and graduate level programs are pending approval. The Elements of AI course is available in Lithuania and accessible to all citizens in Lithuanian, with around 20% of Lithuania's population actively invited to take the course.

3 Technical approaches to disinformation detection

Beyond strategic policies, legislation and educational initiatives, there is an ongoing effort to develop technical solutions for detecting online disinformation, including AI-generated content. This section provides a brief overview of different methodologies for disinformation detection before then developing the statistical approach taken by the SOLARIS project which may be integrated into Use Case 2 (Fenga 2024). For expanded mathematical proofs, please see appendices A and B.

3.1 Disinformation detection methodologies

Among the diverse methodologies used to detect disinformation, two prominent approaches stand out: (i) anomaly detection and (ii) content-based analysis. Anomaly detection is fundamentally anchored on the premise that disinformation often exhibits abnormal patterns when contrasted with the genuine information, and thus this approach is rooted in statistical analysis that seeks to identify patterns that deviate significantly from the norm, particularly in the dissemination of this information. The anomaly detection approach thrives on capturing unexpected shifts, such as abrupt spikes in social media activity or unusual fluctuations in keyword usage, which can be indicative of mis/disinformation campaigns gaining traction among online communities. In contrast, content-based approaches focus on the intrinsic characteristics of the information itself, often leveraging natural language processing (NLP) techniques. This approach scrutinizes linguistic and contextual elements of news articles or social media posts to discern indicators of mis/disinformation. By analyzing word choice, sentiment, and other linguistic patterns, content-based methods can reveal deceptive narratives that might be overlooked by statistical anomaly detection. This emphasis on linguistic analysis underscores the nuanced nature of communication in digital media, where subtleties in language can significantly affect the perceived credibility of information.

Each approach has its unique strengths. Where anomaly detection excels in identifying irregularities in information flow, while content-based analysis provides a nuanced understanding of textual content. Thus integrating both anomaly detection and content-based strategies holds promise for a more robust solution and a more sophisticated model that not only flags potential mis/disinformation effectively but that also provides contextual insights that are vital for understanding the dissemination of disinformation in real time. The convergence of these methodologies can enhance the responsiveness and accuracy of disinformation detection systems, offering a more comprehensive toolkit for combating mis/disinformation.

Balancing resources with availability of tools and researcher expertise, the SOLARIS project has chosen an anomaly detection approach. Ultimately, anomaly detection emerges as a frontrunner in the battle against disinformation due to its statistical rigor, real-time capabilities, and resilience to evolving deceptive strategies. While content-based approaches play an essential role in understanding the intricate linguistic nuances of disinformation, they may falter in contexts where non-linguistic anomalies are prevalent.

3.2 Anomaly detection

A growing body of research focused on detecting mis/disinformation through anomaly detection reveals the effectiveness of leveraging anomaly detection techniques to identify mis/disinformation online. This approach is fundamentally anchored on the premise that the fake news often exhibits abnormal patterns when contrasted with the genuine information. Leveraging on this distinction, it may be possible to develop effective detection mechanisms that can dynamically adapt to evolving mis/disinformation techniques.

Key studies shed light on various aspects of anomaly detection as a methodology. Research by Perez-Rosas, Kleinberg and Shwartz (2018) investigates linguistic features associated with fake news. Their work emphasizes the development of an automated system that utilizes anomaly detection to identify linguistic irregularities, which underscores the significance of such anomalies as reliable indicators of mis/disinformation. This reflects a growing trend in research that combines linguistic analysis with statistical methodologies. Similarly, Ruchansky, Seo and Han (2017) present a comprehensive data analytics frameowkr that incorporates anomaly detection to analyze temporal patterns in the spread of information. Their contributions add a valuable dimension to fake news detection strategies by not only recognizing which content is shared but also discerning the timing and context of its dissemination. This temporal analysis is critical in understanding how narratives spread and which factors drive user engagement. Furthermore, Wang and Zubiaga (2018) amalgamate deep learning with anomaly detection, demonstrating how this synergy can enrich the understanding of irregular patterns in the context of mis/disinformation. Their findings indicate that traditional anomaly detection can be greatly enhanced when paired with advanced machine learning techniques, enabling more accurate predictions and differentiations between genuine and fraudulent content. Adding to this discussion, Popat, Ruhcansky and Mei (2018) introduce an evidence-aware model that integrates deep learning, emphasizing the critical role of anomaly detection in identifying inconsistencies during evidential analysis for debunking fake news. Their model leverages a multi-faceted approach that not only examines the content itself but also evaluates the credibility of the sources contributing to the information, thus reinforcing the importance of context in fake news detection.

A panoramic review by Zubiaga, Li and Ark (2018) highlights the importance of anomaly detection in scrutinizing user behaviour, linguistic nuances, and propagation dynamics. This review positions anomaly detection as a pivotal element in the broader landscape of fake news detection by illustrating how intricate relationships among user interactions, timing, and content characteristics can inform detection strategies. The rationale for this study is underscored by an extensive analysis by Vosoughi, Roy and Aral (2018) which asserts that falsehoods disseminate significantly farther, faster, deeper, and more broadly than the truth across various information categories. Notably, the rapid spread of false information is often attributed to retweets by users rather than automated bots, revealing distinctive patterns of participation on platforms such as Twitter. Studies indicate that false new stories are on average 70% more likely to be retweeted than true stories and reached a wider audience much more quickly. This suggests the need for detection systems to be both proactive and reactive, capable of responding to the immediate dynamics of social media engagement.

3.3 Google Trends as a testing ground for disinformation detection

The SOLARIS approach utilizes data from the Google Trends repository, a free tool that allows users to explore the popularity of search queries over time. Google Trends provides insights into search volume changes for specific terms and visualizes trends and patterns. It assigns a relative score to search terms (ranging from 0 to 100), indicating their popularity during a particular period, making it a valuable resource for analyzing public interest dynamics. While the platform does not provide explicit frequency values, it offers a normalized measure of search interest, facilitating comparisons across different search terms and timeframes. The proposed method leverages Google Trends for the following tasks:

- *Synthesizing Anomalies:* Researchers can create synthetic datasets representing fake news dissemination by introducing anomalies in search interest for keywords associated with misinformation, allowing for robust testing of detection algorithms.
- *Temporal Dynamics Analysis:* The temporal aspect of Google Trends data is essential for evaluating algorithms. They can be tested on their ability to identify irregular patterns, sudden spikes, or fluctuations that indicate the spread of false information.
- *Real-time Testing*: With near-real-time data, researchers can assess algorithm adaptability to emerging trends, ensuring effectiveness in identifying anomalies as search patterns evolve.
- *Baseline Comparison*: By comparing algorithm performance against naturally occurring search baseline trends, the algorithms' ability to distinguish synthetic anomalies from authentic patterns can be validated, enhancing testing reliability.
- *External Factors Consideration*: Google Trends data reflects external influences like major events or shifts in societal interests, allowing researchers to ensure algorithms remain effective amidst real-world variations.

The Google Trends data is updated frequently (every 8 minutes, hourly, daily, weekly, and monthly), enabling monitoring of fake news at various intervals. This multi-frequency approach facilitates real-time insights while providing a broader understanding of information spread over time. The 8-minute interval offers immediate alerts to emerging trends and anomalies, while the 1-hour and 1-day intervals help identify patterns that may not be visible in shorter time frames. Extending to 1-week and 1-month intervals allows the identification of sustained trends and shifts in public interest.

In summary, monitoring fake news across different frequencies, from real-time observation to monthly assessments, provides a comprehensive strategy. Each temporal scale offers unique insights for the detection and mitigation of fake news, enabling a responsive and resilient monitoring system to the challenges of misinformation in the digital landscape.

3.4 Experimental setup

This section describes the experimental design for evaluating fake news detection algorithms using the SOLARIS approach. We use an agent-based model simulating Google Trends activity, combining keyword-based anomaly detection with text analysis.

Modelling Genuine and Fake News

Genuine news dissemination is modelled using relevant keywords tracked across multiple social media platforms (Google Trends, Twitter, Facebook). A keyword's popularity is represented as a time series: $Y_t, t \in T$, where Y_t is the popularity at time t, and T represents the set of discrete time points. The complete time series is $\{Y_1, Y_2, \ldots, Y_T\}$. For multivariate analysis (multiple platforms or keywords), Y_t becomes the following vector:

$$Y_t = \begin{bmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{m,t} \end{bmatrix}$$ where $Y_{i,t}$ is the ith variable at time t.

Fake news (F_t) is modelled as an anomaly (i.e., a significant deviation from the expected pattern μ_t) quantified as $F_t = Y_t - \mu_t$. A large positive or negative F_t value indicates a fake news event. Detection involves comparing $|F_t|$ against a threshold using bootstrapped statistical tests. A time series with fake news introduced at time t_0 is represented as $\{Y_1 - F_1, Y_2 - F_2, \cdots, Y_{t_0} - F_{t_0}, \cdots, Y_T - F_T\}$, where $F_t = 0$ for $t < t_0$ and $F_t \neq 0$ for $t \geq t_0$. This is illustrated in Figure 1.

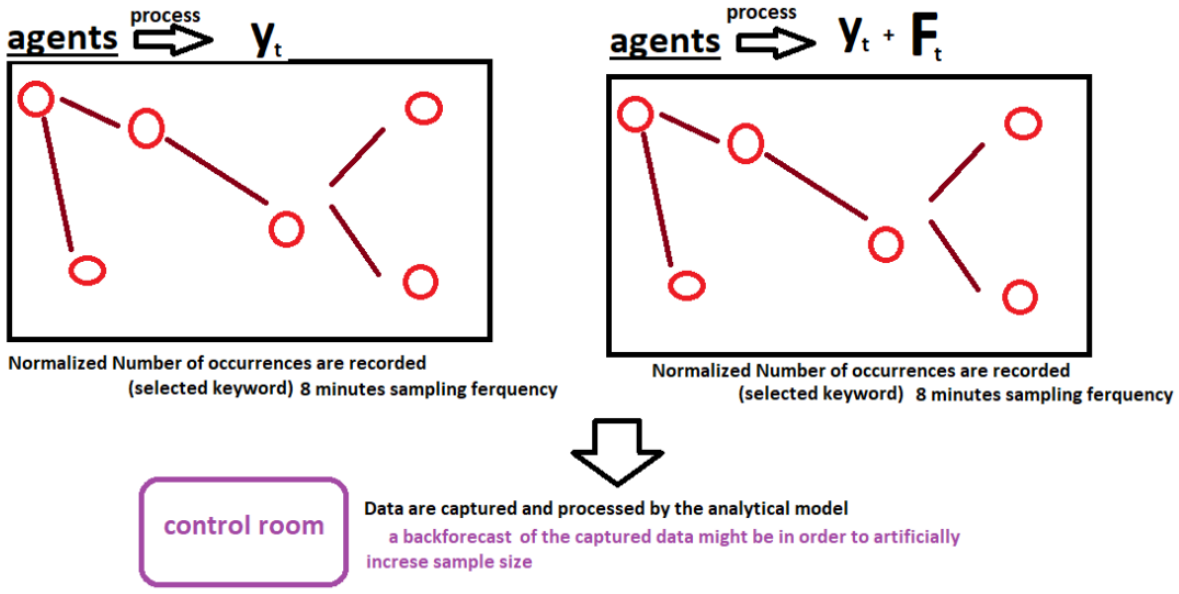

Figure 1. Time series representation genuine and the introduction of fake news at t_0

Agent-based simulation and experiment design

Our agent-based model simulates Google Trends activity using a simplified setup: a single platform (Google Trends), one keyword (Z, representing genuine news with popularity Y_t), and three agents (A, B, C) forming a network (N). The time span is 24 hours with 8-minute sampling. The model incorporates: agent creation with attributes (search behaviour, interest, influence), topic selection (keywords), search behaviour dynamics, interest dynamics, agent interaction, search volume tracking, time evolution, and visualization of search trends. The experiment proceeds as follows:

1. Agents (A, B, C) interact over network N.
2. A genuine news item (R) and keyword (K) are selected.
3. R originates from agent A.
4. A simulated Google Trends environment is used for monitoring.
5. Keyword Y serves as the analysis proxy. A fake news event (F_t) is introduced at t_0, where $F_t = s$ (a constant) for $t \geq t_0$ and $F_t = 0$ otherwise.
6. The algorithm processes $Y_t + F_t$ for $t = 1, \cdots, T$.
7. Algorithm results are recorded.

Advanced scenario: self-correction

In an advanced scenario, a self-awareness mechanism is triggered at $t_0 + 1$, resolving the fake news impact by $t_0 + 3$ (see Figure 2). A stochastic counter-information mechanism is introduced at $t_0 + 3$ to assess the system's self-correcting capabilities. Formally: $F_t = 0$ for $t < t_0$ and $t \geq t_0 + 3$, and $F_t \neq 0$ otherwise.

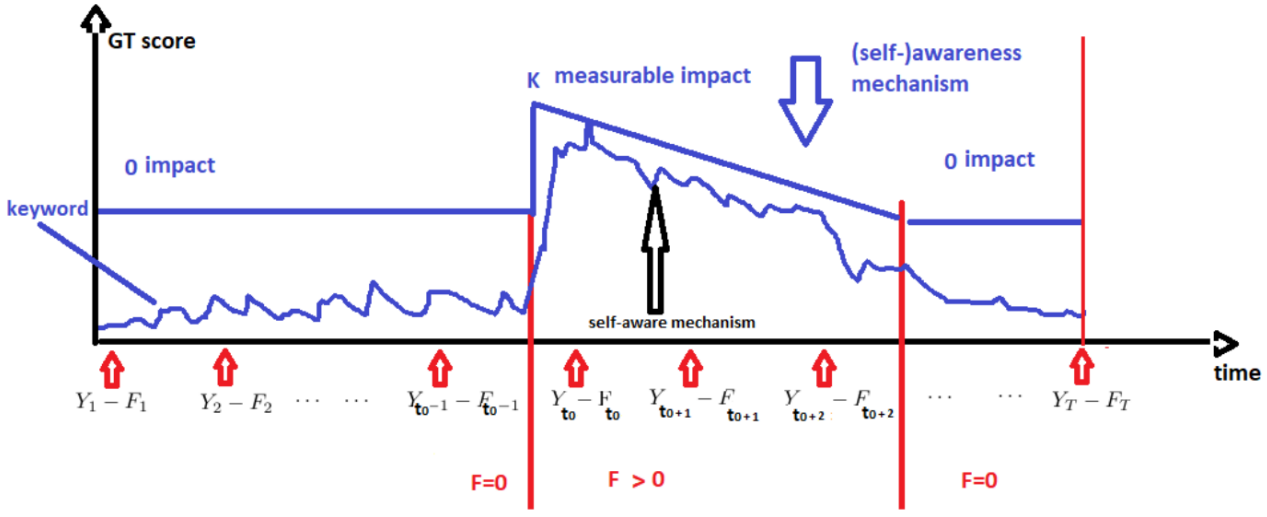

Figure 2. Advanced scenario with self-awareness and counter-information

3.5 Alternative approaches and future work

This study primarily uses keyword-based anomaly detection. Future work will explore text analysis and other platforms to enhance generalizability. Google Trends was chosen for its wide usage, time-series data, and experimental replicability. Further research includes cross-checking fake news across keywords and collaborations for publication.

4 Considerations for Use Case 2

In light of the overviews provided in the previous sections, this section outlines major considerations to be taken forward when developing Use Case 2 for the SOLARIS project. Use Case 2 aims to develop methodologies and policy recommendations for addressing the negative implications of AI-generated content, particularly deepfakes, in order to equip policymakers and media organizations with the necessary strategies to handle this issue. To achieve this objective, Use Case will simulate the dissemination of harmful AI-generated content and the necessary actions taken by media organizations and authoritative institutions in order to mitigate the impact of this content. The actions examined during this simulation will include, for example, the established processes that media organizations and policymakers would follow in order to mitigate the spread of harmful AI-generated content, the types of institutional interactions they would seek to enhance effectiveness, the communication channels already in place for coordinating mitigation strategies, and the main barriers to an effective coordinated response mechanism. A summary of the considerations is provided below in Table 1.

Table 2. Considerations for Use Case 2

Issue	**Considerations for Use Case 2**
Ambiguous definition of AI-generated disinformation (domain, harm, scale)	Observe how stakeholders define disinformation Simulate 2 key scenarios: (i) “natural” spread and (ii) coordinated dissemination
Public concern for freedom of speech	Observe how public concerns impact decision-making Observe how stakeholder’s view their own actions in countering disinformation.
Technical and organizational obstructions	Include local/region-specific stakeholders Assess differences in stakeholder decision-making for incidents occurring in different regions
Coordination between organizations	Include diverse stakeholders and observe how responses align/diverge
Quantitative measures for early detection	Experiment with indirect indicators for monitoring the spread of harmful disinformation
Effectiveness of response protocols	Assess how/if pre-determined protocols are followed Assess effectiveness of protocols in real-time scenarios
Impact and use of media literacy / fact checkers	Assess how stakeholders gauge audience literacy needs and promote media literacy Assess how media literacy informs public reporting

4.1 Defining the problem of AI-generated disinformation

While it is widely recognised that generative AI technologies will greatly improve the quality and scale of disinformation campaigns, the specific issue of AI-generated disinformation is ill-defined across legislation, policy and strategy in the current European landscape. In the context of AI governance legislation and policy, such as in the AI Act and in various national AI strategies, the issue of AI-generated disinformation is often recognised as one of several motivators for developing more robust AI regulation. However, it is often sidelined as a minor or peripheral issue in proposed solutions while other socio-political issues such as algorithmic bias, surveillance and labour impact remain the focus. In the context of counter-disinformation strategies and cybersecurity, AI-generated disinformation is similarly overlooked or often simply equated to traditional disinformation and it is often assumed that current tactics can be simply extended to AI-generated disinformation. As such, there are little to no explicit policies or strategies that are aimed at directly addressing the issue of AI-generated disinformation as a distinct problem requiring new and specific responses, as many experts have called for. It is worth noting that this is not simply an issue for EU Member States as AI-generated disinformation is similarly ill-defined in UK policy and legislation.

The ambiguity around the issue of AI-generated disinformation further extends to how the problem is conceptualized more broadly. In terms of scale, disinformation can be understood as a problem in which harmful individual content is spread "naturally" between online users and thus requires more robust moderation mechanisms to identify or censor such content, as is the approach of the UK's Online Safety. However, disinformation can also be conceptualized on a much larger scale as coordinated and motivated campaigns (often instigated by foreign actors) involving the spread of harmful narratives through multiple pieces of online content and that thus requires a coordinated national response, as is approach outlined in national strategies of countries such as Spain and France. Furthermore, much of the literature regarding disinformation and AI-generated disinformation is narrowly focused on electoral interference and the actions of foreign actors. Certainly, elections constitute flashpoints where political manipulation is heightened and where foreign adversaries can more readily manipulate the information environment to cause harm. However, this focus on elections risks overlooking the continuous role that disinformation plays in encouraging polarization between communities and eroding confidence in democratic processes, government institutions and media organizations. Furthermore, this focus on foreign actors risks overlooking threats from internal actors and individuals, particularly as generative AI technologies have significantly developed in quality and are easily available to the public. These conceptualizations are, of course, not exclusive and often national governments seek to address the problem in different ways depending on the policy area.

With these different conceptualizations of the issue itself comes further ambiguity around what constitutes harmful fake content. Notably, the UK identifies harmful content as that which causes psychological or physical harm upon an individual, for example the distress caused by deepfake pornography. In contrast, the DSA considers the broader societal harms of disinformation and other national criminal codes such as those in Italy, Spain and Albania characterize harm in terms of public order and citizen safety. This

broader understanding of societal harm rather than personal harms also accounts for hate speech, defamation and slander.

In light of this ambiguity, Use Case 2 should consider and observe how key stakeholders initially define the problem of AI-generated disinformation and how this initial conceptualization then goes onto influence their decision-making and what institutions they choose to interact with when coordinating a response. Particularly, this should take into account how stakeholders characterize AI-generated disinformation as an issue of AI governance, national security, public order, cybersecurity, or a combination of these, and also how they identify the potential harms during the simulation. This would enable SOLARIS to better understand whether or not it is necessary to introduce legislation and/or policy that is specific to AI-generated disinformation. Additionally, Use Case 2 should consider simulating both (i) a situation in which harmful individual content spreads online naturally and (ii) a coordinated and motivated campaign to spread multiple pieces of harmful content online in order to promote a specific narrative. While it would be beneficial to pursue a much longer simulation in order to consider how AI-generated disinformation contributes to the long-term erosion of trust in democracy, this is not viable within Use Case 2.

4.2 Challenges to implementation

While there are strategies and legislation in place to tackle the spread of AI-generated disinformation (albeit ill-defined), there remain issues with implementation that differ significantly between Member States, regions and policy areas. Across all social domains, however, there is a fundamental concern that any form of counter-disinformation strategy may amount to government censorship and thus infringing upon people's right to freedom of speech. This concern has been raised as a criticism in response to almost every piece of legislation intended to tackle disinformation, a notable example being the Avia Law in France. While there are legitimate concerns about providing governments, media organizations and large technology companies with the power to dictate what constitutes appropriate information, they also present an obstacle to mitigating the impacts of disinformation. These anxieties are only further exacerbated by the ambiguity over definitions of what constitutes harmful content (as discussed above) that risk being interpreted selectively and also by the current political landscape that is increasingly polarized and distrustful of government institutions and media organizations, a situation that has emerged in part as a result of online disinformation. Within this current climate of uncertainty, stricter counter-disinformation policies to tackle AI-generated disinformation may be received negatively by the public. Use Case 2 should take into account this public sentiment and potential backlash by observing whether or not stakeholders feel that the actions they take in order to tackle AI-generated disinformation are inappropriate or amount to a form of censorship. It should further consider what decisions they make or concerns they weigh up before taking any action against AI-generated disinformation and particularly whether or not they are afraid that they may be accused of censorship and if this fear influences their actions.

Beyond issues of public support of counter-disinformation strategies, there are other more practical issues obstructing the implementation of these strategies. Where some Member States have highly developed

government institutions, country-wide organizational networks, substantial funding, and the technical infrastructure necessary for implementation, this is not the case for other smaller Member States such as those in the Western Balkans. While this regional scale is beyond the scope of Use Case 2, the simulation could include local or region-specific stakeholders and it could seek to assess what different steps stakeholders take in order to address the spread of disinformation in different regions, if any. This would allow SOLARIS to better understand how the effectiveness of stakeholder decision-making is impacted by regional differences and whether limitations (e.g., technical, organizational infrastructure) change their tactics.

Even were the technical and organizational infrastructures in place, implementing counter-disinformation strategies may be further obstructed by coordination and cooperation between organizations at the local, national, and European level. Organizations at different levels will operate according to different policies, exercise different jurisdictions, and have different resources at their disposal such that coordination between these organizations can be difficult. This is a notable issue for local journalism which has declined significantly in recent decades. Some countries have sought to address this issue, notably Spain introduced the Protocol to Combat Disinformation that emphasizes inter-agency cooperation while the UK has introduced regional cybersecurity hubs to coordinate responses. However, many other States suffer from a lack of coordination. In Use Case 2, it is necessary to understand how different organizations respond to AI-generated disinformation, how their resources and relationships to the impacted community influence their actions, and how their different responses harmonise or conflict with one another. This could be achieved by including more diverse stakeholders within the simulations, notably local journalists and local government officials alongside national and international stakeholders.

A final significant obstacle to implementing counter-disinformation strategies is the increasingly fragmented media landscape including traditional news media and social media platforms, both across Europe and within Member States. Due to the widespread availability of digital technologies and the proliferation of social media, there has been a drastic increase in the number of people producing and/or disseminating information online. As such, many online users and entire communities do not share common sources of information and people increasingly receive highly personalized media via recommendation algorithms. This explosion of online news sources and media platforms means that it is difficult to ensure that monitor online information and to ensure that these sources abide by counter-disinformation legislation. Notably, many of the provisions of the DSA only apply to very large online companies such that smaller but influential sources may be overlooked. Meanwhile, soft law tools such as self-regulation introduced by the EU have often proved ineffective as they voluntary and it is difficult to monitor the activities of these platforms. In order to account for this increase in information sources, Use Case 2 should simulate the spread of AI-generated disinformation on various different social media platforms and online news sources, particularly smaller ones.

4.3 Preparedness and prevention

While the above has focused on conceptualizing the problem of AI-generated disinformation and of the challenges to counter-disinformation responses, it is also necessary to consider methods for prevention and

preparedness. Currently, there are numerous available methods for detecting disinformation that rely upon automated technologies to identify and moderate possibly harmful content spreading online. However, these methods can feature inaccuracies that may result in genuine content being flagged as disinformation, further contributing to concerns over limiting free speech. SOLARIS uses a detection method that monitors the changes in searches on Google over time and identifies anomalous activity as a potential indirect indicator of disinformation. For example, a sudden increase in users searching the term "nuclear" may indicate that disinformation featuring nuclear weapons has begun to spread online. This anomaly detection method may be integrated into Use Case 2 but it further raises the prospect of using other indirect indicators to monitor for the emergence of harmful disinformation. In Use Case 2, stakeholders from media organizations could be provided with a range of relevant simulated quantitative data (e.g., social media behaviour, audience interaction, article views) and their reactions could be monitored in order to determine which if any of these indicators could be used to indirectly monitor disinformation spreading.

With regards to preparedness, many government cybersecurity and communications authorities, such as the GCHQ in the UK, have issued guidance for private businesses and public organizations on how to tackle a disinformation threat. Given the diversity and ambiguity around disinformation threats, it remains unclear how effective this guidance is in a real-time scenario. Use Case 2 provides an opportunity to better understand these protocols and how they are used by individual stakeholders to appropriately tackle a disinformation threat or escalate the problem to higher authorities or possibly national security organizations. To address this, Use Case 2 should seek to understand what precise escalation protocols are in place (if any), whether or not stakeholders are actively aware of these protocols, if or how they carry out these protocols in real time, and whether or not such protocols are sufficient to address a disinformation threat.

Aside from technical detection methods and organizational procedures, there has been a significant drive toward improving media literacy at both a national and European level. However, there is a lack of initiatives focusing on AI literacy specifically and often these campaigns are limited in their reach beyond young people in schools. Media literacy campaigns need to continue to evolve in order to respond to the changing technological and political landscape. Furthermore, education initiatives beyond media literacy need to address the root causes such as community polarization and the psychological reasons for false belief systems. However, these efforts may be seen as a form of manipulation and thus may have an adverse effect causing further distrust in institutions. While Use Case 2 is primarily focused on response, media literacy concerns may be taken into account by asking stakeholders how they gauge their audience's literacy needs, as well as what their organization has done to promote media literacy (e.g., fact-checking initiatives) and if such efforts are effective. Following the simulated scenario in Use Case 2, it may also be insightful to discuss with stakeholders how they publicly report such an incident in a way that promotes literacy and whether or not they would alter their literacy efforts in response to this specific incident of AI-generated disinformation.

References

Artificial Intelligence Act, EU Parliament legislative resolution of 13th of March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM (2021) 0206-C9-0146/2021-2021/0106 (COD)).

AI Governance Alliance (2024). *Generative AI Governance: Shaping a Collective Global Future*. World Economic Forum.

Ajji, Kamel (2020) "Protecting liberal democracy from artificial information: the French proposal" in *Fundamental Rights Protection Online*. Eds. Petkova, Bilyana and Ojanen, Tuomas. Edward Elgar Publishing: 57-83.

Belli, Luca (2020) *CyberBRICS: Cybersecurity Regulations in the BRICS Countries*. Cham, Switzerland: Springer Nature.

Bieber, Florian and Kmezic, Marko (2015). *Media Freedom in the Western Balkans*. Balkans in Europe Policy Advisory Group.

Bjeloš, Maja and Pavlović, Marija (2022) *Serbia: Drawing the Links to Human Rights and Investing in People*. Belgrade Centre for Security Policy.

Blocman, Amelie (2019) "Laws to Combat Manipulation of Information Finally Adopted". *European Audiovisual Observatory*.

Clark, Adam (2024) *Cybersecurity in the UK*. House of Commons Library.

Create Lithuania (2019) *Lithuanian Artificial Intelligence Strategy: A Vision of the Future*. Ministry of the Economy and Innovation.

DCMS (2021) *Online Media Literacy Strategy*. UK Government Department for Digital, Culture, Media and Sport.

De Witte, Bruno (2023) "Soft law in European public law" in *Research Handbook on Soft Law*. Eds. Eliantonio, Mariolina and Korkea-aho, Emilia. Cheltenham, UK: Edward Elgar Publishing.

Draghi, Mario (2024) *The Future of European Competitiveness*. European Commission.

Dunn, Pietro (2024) *Online Hate Speech and Intermediary Liability in the Age of Algorithmic Moderation*. Doctoral Thesis, University of Luxembourg.

Estonian Government (2019) *Estonia's National Artificial Intelligence Strategy 2019-2021*. Ministry of Economic Affairs and Communications.

EU Disinfo Lab (2021) "Why Disinformation is a Cybersecurity Threat". *EU Disinfo Lab*. 24 May 2021. <https://www.disinfo.eu/advocacy/why-disinformation-is-a-cybersecurity-threat/#:~:text=Placing%20disinformation%20policy%20at%20the%20heart%20of%20the,a%20robust%20civil%20society%20and%20more%20resilient%20EU.>

European Commission (2001) *European Governance White Paper*.

European Commission (2003) *Enhancing the Implementation of the New Approach Directives*.

European Commission (2016a) *European Strategy on Countering Hybrid Threats*.

European Commission (2016b) *General Data Protection Regulation*.

European Commission (2018a) *Audiovisual Media Services Directive*.

European Commission (2018b) *Code of Practice on Disinformation*.

European Commission (2018c) *Tackling Online Disinformation: A European Approach*.

European Commission (2019a) *European Media Literacy Week*.

European Commission (2019b) *Cybersecurity Act*.

European Commission (2020a) *European Democracy Action Plan*.

European Commission (2020b) *EU's Cybersecurity Strategy for the Digital Decade*.

European Commission (2020c) *Tackling Covid-19 Disinformation*.

European Commission (2021a) *Digital Education Action Plan 2021-2027*.

European Commission (2021b) *Guidance on Strengthening the Code of Practice on Disinformation*.

European Commission (2022a) *Guidelines for teachers and educators on tackling disinformation and promoting digital literacy through education and training*.

European Commission (2022b) *Digital Services Act*.

European Commission (2022c) *NIS2 Directive*.

European Commission (2022d) *Proposal for Regulation of the European Parliament and of the Council on Horizontal Cybersecurity Requirements for Products with Digital Elements*.

European Commission (2022e) *Report on Building Resilience Against Disinformation*.

European Commission (2022f) *Strengthened Code of Practice on Disinformation*.

European Commission (2024) *European Digital Identity Framework*.

European Parliament, Council and the Commission, 2003, Interinstitutional Agreement on Better Lawmaking, OJ C 321/01.

Farrand, Benjamin (2024) "How Do We Understand Online Harms? The Impact of Conceptual Divides on Regulatory Divergence Between the Online Safety Act and Digital Services Act" *Journal of Media Law*.

Fenga, Livio (2024) "A Hybrid Comuter-Intensive Approach Integrating Machine Learning and Statistical Methods for Fake News Detection" *Engineering Proceedings* 68.47.

French Mission Report (2019) *Creating a French framework to make social media platforms more accountable: acting in France with a European vision*.

Full Fact (2024) *Trust and Truth in the Age of AI*. Full Fact Report 2024.

Gayo, Miguel Recio (2019) *Protección de datos personales e innovación: ¿(in)compatibles?*. Madrid, Spain: Editorial Reus.

GCS (2022a) *RESIST 2: Counter-disinformation Toolkit*. UK Government Communication Service.

GCS (2022b) *The Wall of Beliefs: A Toolkit for Understanding False Beliefs and Developing Effective Counter-disinformation Strategies*. UK Government Communication Service.

General Guidelines for the Cooperation between CEN, Cenelec and ETSI and the European Commission and the European Free Trade Association, 1984.

Gobierno de España (2019). *National Counter-Terrorism Strategy 2019.*

Gobierno de España (2021) *Estrategia de Seguridad Nacional 2021*.

Górka, Marke (2023). "Baltic States Cyber Security Policy: Development of Digital Capabilities in 2017-2022". *Stosunki Międzynarodowe – International Relations*.

Implement (2023) *The Economic Opportunity of Generative AI in D9+*. Implement Consulting Group.

Judson, Ellen, Kira, Beatriz, and Howard, Jeffrey W (2024) "The Bypass Strategy: Platforms, the Online Safety Act and Future of Online Speech" *Journal of Media Law*.

Kapo, Amra (2023). *Mapping of actors: identification of potential cooperation partners for media outlets in the Western Balkans*. GIZ GmbH.

Kira, Beatriz (2024) "When non-consensual intimate deepfakes go viral: the insufficiency of the UK Online Safety Act" *Computer Law & Security Review: The International Journal of Technology Law and Practice* 54.

Kokalari, Amy (2024) *Digitalized Western Balkans: Asset or Liability to EU Enlargement*? Master's Thesis, European University Institute.

Krivokapić, Djordje, Živković, Ivona, and Nikolić, Andrea (2022) "Artificial intelligence regulation in the areas of data protection, information security, and anti-discrimination in Western Balkan economies". *2022 45th Jubilee Internation Convention on Information, Communication and Electronic Technology*.

Latvian Government (2020) *Developing Artificial Intelligence Solutions*. Latvian Government.

Legrand, Emmanuel (2021) "France's broadcasting law creates a new regulator for audiovisual and digital communication. *Creative Industries News*. 8 November 2021. <https://creativeindustriesnews.com/2021/11/frances-broadcasting-law-creates-a-new-regulator-for-audiovisual-and-digital-communication/>

Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. (2018). Boletín Oficial del Estado.

Loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information. (2018). Journal officiel de la République française.

Margova, Ruslana & Dobreva, Milena (2023) *Disinformation Landscape in Bulgaria*. EU Disinfo Lab.

Marotta, Manny (2020) "France Constitutional Court strikes down most of online hate speech law". *Jurist News*. 20 June 2020.

Montasari, Reza (2023) *Countering Cyberterrorism: The Confluence of Artificial Intelligence, Cyber Forensics and Digital Policing in US and UK National Cybersecurity*. Cham, Switzerland: Springer Nature.

Mündges, Stephen and Park, Kirsty (2024) "But did they really? Platforms' compliance with the code of practice on disinformation in review". *Internet Policy Review* 13.3: 15-16.

Nannini, Luca, Bonel, Eleonora, Bassi, David, Maggini, Michele Joshua. "Beyond phase-in: assessing impacts on disinformation of the EU Digital Services Act". *AI Ethics*.

Nenadic, Iva, Brogi, Elda, Bleyer-Simon, Konrad (2023) *Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation*. Centre for Media Pluralism and Media Freedom, European University Institute.

Nenovski, Blagoj, Ilijevski, Ice, and Stanojoska, Angelina (2023) *Strengthening Resilience Against Deepfakes as Disinformation Threats*. University St. Kliment Ohridski Bitola.

NIST (2024) *AI Risk Management Framework*. National Institute of Standards and Technology.

Novelli, Claudio, Hacker, Philipp, Morley, Jessica, Trondal, Jarle, and Floridi, Luciano (2024) "A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities". *European Journal of Risk Regulation*.

OECD (2023) *G7 Hiroshima Process on Generative AI: Towards a G7 Common Understanding on Generative AI*. Organization for Economic Co-operation and Development.

Osborne Clark (2021) "Reform of France's audiovisual sector takes major step forward". *Osborne Clark*. 28 April 2021. https://www.osborneclarke.com/insights/reform-frances-audiovisual-sector-takes-major-step-forward

Perez-Rosas, V., Kleinberg, J., & Schwartz, H. (2018) "Detecting Fake News in Social Media" in *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 550-561). Association for Computational Linguistics.

Popat, K., Ruchansky, N., & Mei, Q. (2018). "A Novel Approach to Detecting Fake News with Evidence" in *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 343-351). Association for Computational Linguistics.

PSOE (2023) "Santos Cerdán anuncia que el PSOE creará un comité para desmentir 'bulos' de la derecha". *PSOE*.

Republic of Albania (2020) *National Cybersecurity Strategy 2020-2025*.

Republic of Albania (2015) *Policy Paper on Cybersecurity 2015-2017*.

Ricci, Alexander Damiano (2018) "France moves to fight the 'manipulation of information' instead of 'fake news'". *Poynter*. 6 June 2018.

Rubinstein, Ira (2018) "The Future of Self-Regulation is Co-Regulation" in *The Cambridge Handbook of Consumer Privacy*. Eds. Selinger, Evan, Polonetsky, Jules, and Tene, Omer. Cambridge, UK: Cambridge University Press.

Ruchansky, N., Seo, S., & Han, S. (2017). "Cassandra: Detecting Fake News by Using Natural Language Processing and the Social Web" in *Proceedings of the International Conference on Data Science* (pp. 781-789). IEEE.

Schepel, Harm (2005) *The Constitution of Private Governance: Product Standards in the Regulation of Integrating Markets*. London, UK: Hart Publishing.

Senden, Linda (2004) *Soft Law in European Community Law*. Oxford, UK: Hart Publishing.

Senden, Linda (2005) "Soft Law, Self-Regulation and Co-Regulation in European Law: Where Do They Meet?". *Electronic Journal of Comparative Law* 9.1.

Senden, Linda et al. (2015) "Mapping Self- and Co-regulation Approaches in the EU Context". *Explorative Study for the European Commission*.

Shu, K., Sliva, A., Wang, S., & Lee, D. (2019). "Fake News Detection on Social Media: A Data Mining Perspective". *ACM SIGKDD Explorations Newsletter* 21.2: 22-36.

Staff Working Document (2020) *Assessment of the Code of Practice on Disinformation*.

Stasi, Maria Luisa and Parcu, Pier Luigi (2021). "Disinformation and Misinformation: the EU Response" in *Research Handbook on EU Media Law and Policy*. Eds. Pier Luigi Parcu and Elda Brogi. Cheltenham, UK: Edward Elgar Publishing.

Tsamados, Andreas (2021) *Artificial Intelligence for Social Good: Avoiding the Solutionist Trap*. Hellenic Foundation for European & Foreign Policy.

Tufa, Alban (2023) "Avantazhet dhe disavantazhet e inteligjencës artificiale në Shqipëri dhe masat e ndërmarra nga institucione përkatëse dhe Bashkimi European" *EuroSpeak*. 22 September 2023. <https://www.eurospeak.al/news/fuqia-e-bute/ai-chatgpt-akademia/>

UK Government (2021) *National AI Strategy*.

UK Government (2022) *Government Cyber Security Strategy: Building a Cyber Resilient Public Sector*.

UK Government (2023) *Online Safety Act*.

UNDP (2023) *Digital Readiness Assessment: North Macedonia*. United Nations Development Programme.

UNESCO (2024) *Generative AI: Navigating the Opportunities and Risks of Artificial Intelligence in Education*.

Vasileva, Konstantina (2023) "News media as part of the disinformation landscape". *AEJ Bulgaria*. 8 February 2023. https://aej-bulgaria.org/en/news-media-as-part-of-the-disinformation-landscape/

Vicente, Ana Romero (2023) "The Disinformation Landscape in Spain". *EU Disinfo Lab*. 1 March 2023.

Villasenor, John (2019) "Artificial intelligence, deepfakes, and the uncertain future of truth". *Brookings Institute*. 14 February 2019.

Vosoughi, S., Roy, D., & Aral, S. (2018). "The spread of true and false news online", *Science,* 359.6380: 1146-1151.

Wachter, Sandra (2024) "Limitations and Loopholes in the EU AI Act and AI Liability Directives: What this means for the European Union, the United States and Beyond". *Yale Journal of Law and Technology* 26.3.

Wang, S. & Zubiaga, A. (2018). "A Comprehensive Study on Detecting Fake News on Social Media" in *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 140-145). Association for Computational Linguistics.

Weidenger, Laura et al. (2022) "Taxonomy of Risks Posed by Language Models". *Proceedings of the 2022 ACM Conference on Fairness, Accountability and Transparency*.

WEF (2024) *AI Governance Alliance*. World Economic Forum.

Wihbey, John (2024) "AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?". Northeastern University Ethics Institute.

Young, Zachary (2018) "French Parliament passes law against 'fake news'". *Politico*. 4 July 2018.

Zubiaga, A., Li, H., & Ark, G. (2018) “A Survey on the Role of Social Media in Fake News Detection”. *Journal of Data Science* 16.2: 245-267.

Appendix A. Methodology

A1. ARIMA models

ARIMA models capture temporal dependencies and trends in time series data. A non-seasonal ARIMA model, denoted as ARIMA(*p, d, q*), consists of three components:

- An autoregressive (AR) component of order p : $\sum_{i=1}^{p} \emptyset_i Y_{t-i}$
- An integrated (I) component of order d, representing differencing to achieve stationarity: $\nabla^d Y_t$
- A moving average (MA) component of order q : $\sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$

The general ARIMA(*p, d, q*) model is expressed as:

$$\nabla^d Y_t = \sum_{i=1}^{p} \emptyset_i \nabla^d Y_{t-1} + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

where Y_t is the observed value at time *t*, ε_t is the white noise error term, $\emptyset_i$ are the AR coefficients, and θ_i are the MA coefficients. The differencing operator ∇ is defined as $\nabla Y_t = Y_t - Y_{t-1}$.

A2. Exponential smoothing models

Exponential smoothing models assign exponentially decreasing weights to past observations, giving higher weight to more recent data. Simple exponential smoothing, suitable for data without trend or seasonality, is defined as: $\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t$, where $\hat{Y}_{t+1}$is the one-step-ahead forecast, Y_t is the observed value at time *t*, $\hat{Y}_t$ is the one-step-ahead forecast at time *t*, and $\alpha (0 < \alpha < 1)$ is the smoothing parameter. More complex models, such as Holt-Winters, incorporate trend and seasonal components for improving forecasting accuracy.

Appendix B. The Algorithm

The code provided is written in the R programming language and initially focuses on importing and preprocessing data for time series analysis. It begins by loading several R packages (libraries) that will be used in the analysis. These packages are essential for various data manipulation, time series modelling, and visualization procedures. The data are then imported from a CSV file located at the specified file path and transformed into a time series object with a frequency of "12", meaning that the data are representative of monthly observations.

B1. Definition of the training and test sets

The code proceeds to define the in-sample and out-of-sample sets for the time series analysis, calculated as follows:

- the training set is the whole time series which of length, say T except for the last k+1 observations.
- the test set begins at T-k

In summary, this section of code is dedicated to importing time series data from a CSV file, transforming it into a time series object, and dividing it into training and test sets. This is a fundamental step in time series analysis, as it prepares the data for subsequent modelling, forecasting, and analysis.

B2. Forecasting

Here, a series of tasks related to forecasting and error evaluation are performed for three different models: Exponential Smoothing (ETS), AutoRegressive Integrated Moving Average (ARIMA), and Extreme Learning Machine (ELM). These tasks are performed over a given number of replications, which is set to '150'.

Bootstrap and Model Fitting Loop

The specified number of replications, which represents the number of resampling iterations to be performed in the subsequent analysis, is assigned to an appropriate variable. This variable determines how many times the subsequent prediction and error evaluation procedures are repeated. At this point, the code enters a 'for' loop, sequentially iterating through each of the 150 replications. For each replication, the following actions occur:

- A bootstrap sample is generated by resampling from the training dataset. This results in a single bootstrap sample that reflects a resampled representation of the training data.
- An ETS model is fit to the generated bootstrap sample, with the model specification set to "ZZZ".
- An ARIMA model is fitted to the bootstrap sample.
- An ELM model is fitted to the bootstrap sample.

Prediction and Error Evaluation

After model fitting, the code proceeds to compute predictions and evaluate errors for each model. It iterates through the prediction horizon, which extends from 1 to the length of the out-of-sample set (12 steps forward).

For each step within this horizon, rolling forecasts are computed for each model, specifically for the time step at horizon t+1. In addition, the mean forecast value for each model at the given time step is extracted and stored in the corresponding vector.

After computing the predictions, the code calculates the forecast errors and percentage errors for each model at each time step. The forecast errors represent the differences between 7 the predicted values and the actual test data. The prediction percentage errors are derived by calculating the absolute values of the prediction errors, dividing them by the test data, and then multiplying the result by 100 to express the errors in percentage terms. For each model and replication, both types of error are recorded in their respective matrices.. In these matrices, each row corresponds to a replication, while each column corresponds to a specific time step. For each replication, the errors are documented in the corresponding row of the matrices. This approach simplifies the storage of errors from multiple replications.

In summary, this code section is responsible for generating rolling forecasts for ETS, ARIMA, and ELM models over a specified number of replications. It calculates the prediction errors and stores them in matrices for subsequent analysis and evaluation. This process is repeated for each model and replication, providing insight into the performance of the forecasting models.

B3. Confidence Interval for Anomaly Detection

This code segment is dedicated to the calculation of intra-interval counts based on the standard deviation for each of the three prediction models: ETS, ARIMA, and ELM. These counts quantify the number of bootstrapped predictions that fall within a predetermined range around the standard deviation of the test set. This process is carried out in order to establish a precise confidence interval for the predictions. This allows the algorithm to detect anomalies within the analysed time series if a significant proportion of the predictions from the 150 bootstrap replications deviate from the actual value by more than two standard deviations.

The first step is to define and assign a value to the parameter 'a', which is fixed at 2. This value significantly influences the width of the confidence interval, covering a range of ±'a' standard deviations around the actual values in the time series data. Next, the code calculates the value of the standard deviation for the test data, a key factor in defining the parameters of the confidence interval. Thus, the within-interval counts are computed for each of the three prediction models. To achieve this, a custom function is applied to each column representing different time steps within the prediction error matrix specific to each model. This function is designed to calculate the number of errors that fall within the predefined confidence interval. Essentially, for each column corresponding to a particular time step, it accumulates the number of errors that satisfy certain Boolean conditions, checking that each error is greater than or equal to '-a * sd value' and less than or equal to 'a * sd value'. Upon completion of this calculation, three variables were successfully derived:

1. The variable containing within-interval counts for the ETS model at each time step.
2. The variable holding within-interval counts for the ARIMA model at each time step.
3. The variable containing within-interval counts for the ELM model at each time step.

In addition to providing valuable insight into the accuracy and reliability of predictive models, these variables are integrated into the subsequent anomaly detection process, as discussed in the following sections.

B4. Statistical Tests

Binomial test

At this point, the assumption of a binomial distribution of the data is introduced in this framework. It is well known that a binomial random variable, symbolically represented as X Binomial ($\pi;n$), characterizes the number of successful outcomes (x) within a series of n independent Bernoullian trials, where the probability of success π remains constant. The binomial probability function can be precisely expressed as follows:

$$P(x) = \binom{n}{k}\pi^x(1-\pi)^{n-x} \text{ for } x = 0, 1, 2, \cdots, n \text{ and } 0 < \pi < 1$$

This code section is dedicated to the execution of binomial tests for each of the three forecasting models. The overall goal of these tests is to determine whether the observed within-interval counts of the forecast errors differ significantly from the expected counts, thereby indicating anomalies within the time series. The code also calculates p-values, determines critical values, and makes decisions about accepting or rejecting the null hypothesis.

Probability calculation

For each of the three models, the code calculates the probability 'π', which represents the probability that the number of successes for the prediction contained in the previously defined confidence interval will fall below the 25th percentile, indicating a 0.05 level of significance. This probability is obtained by calculating the mean of the number of successes (X) within the confidence interval and then dividing by the total number of bootstrap replicates (R):

$$\pi = \frac{mean\ (X)}{R}$$

Critical values

Thus, critical values ($\bar{x}$) are computed for the binomial distribution associated with each model. These critical values indicate the minimum number of successes within the specified interval (25th percentile) before the null hypothesis is rejected. The binomial distribution function is used to derive these values and the results are stored in variables. The code identifies the index corresponding to the integer number of successes within the specified interval that are closest to the significance level. This number is then subtracted from the total number of bootstrap replicates to obtain the threshold for defining the null hypothesis. The results, including the index and the corresponding critical value, are presented for each model. This information is used to indicate the minimum number of successes within the confidence interval to fail to reject the null hypothesis.

Hypothesis testing

The hypothesis test is structured as follows:

$$H_0 : x_{model} < \bar{x}$$
$$H_1 : x_{model} \geq \bar{x}$$

To perform hypothesis testing, the code initializes vectors designed to capture the results of the hypothesis tests for each model. The code then iterates through each time step within the data, representing within-interval counts for each model. At each time step:

- The observed number of successes specific to the model is retrieved.
- A hypothesis test is performed by comparing the observed count with the critical value.
- If the observed count falls below the critical value, the null hypothesis is rejected; otherwise, it is not rejected.

The code reports the results of the hypothesis tests for each model, indicating whether the null hypothesis is accepted or rejected at each time step, and thus whether or not an anomaly has been detected. In summary, the primary objective of this code is to determine whether the forecast series for each model exhibit notable deviations from the expected distribution, as defined by the 25th percentile. The critical values and the results of the hypothesis tests together facilitate the process of detecting anomalies in the time series.