Strengthening democratic engagement through value-based generative adversarial networks

D6.2 SOLARIS

Regulatory innovations and policy options for generative AI and digital democracy GANs

Lead Author: Yasaman Yousefi

With contributions from: Andrew McIntyre, Federica Russo, Maria Dolores Sanchez Galera, Tommaso Tonello

Deliverable nature:	R – Document
Dissemination level: (Confidentiality)	PU
Delivery date:	31/01/2026
Version:	1.3
Total number of pages:	45
Keywords:	Generative Artificial Intelligence, Synthetic Media and Deepfakes, Digital Democracy, Media and AI Literacy, Democratic Resilience, Policy Recommendations

This deliverable has been submitted but not yet approved by the European Commission.

Executive summary

This deliverable (D6.2) explores how synthetic media, particularly audiovisual content generated through various Generative AI systems, influences democratic engagement, reshapes political discourse, and transforms citizens' participation in the digital public sphere.

It identifies **regulatory innovations suitable for governing AI-generated content** in line with the EU's principles of human dignity, accountability, and transparency, in line with those principles, enshrined in the EU Treaties, of human dignity, accountability, and transparency; proposing policy options that empower governments, platforms, and civil society to strengthen digital democracy; and consolidating best practices derived from co-creation methodologies and participatory design experiments.

D6.2 reports on the SOLARIS synthesis book *Deepfakes, Democracy, and the Ethics of Synthetic Media*, particularly Chapters 6 and 8, which examine the ethical, political, and regulatory dimensions of synthetic media. It also integrates insights from relevant EU policy frameworks, such as the AI Act and the Digital Services Act. Building on the SOLARIS network approach, this deliverable distinguishes three principal nodes for coordinated action in governing synthetic media: (i) technology developers and social media platforms, (ii) the regulatory institutions, and (iii) the public as informed and engaged actors within the digital ecosystem.

The findings demonstrate that **transparency and traceability** must become integral features of GenAI systems. This requires the implementation of clear labelling and provenance mechanisms that allow citizens to identify AI-generated content confidently. Furthermore, media literacy, understood as a civic capacity to interpret, contextualise, and ethically evaluate synthetic media, should be promoted as a vital mechanism for ensuring that public values remain embedded throughout the development and dissemination of AI-generated material. Regulatory innovation should combine risk-based oversight with adaptive co-regulation to ensure both flexibility and protection of fundamental rights.

In parallel, digital citizenship education must evolve to include comprehensive AI literacy that complements and reinforces participatory co-creation. Citizens should be equipped to recognise and critically assess synthetic content. They should also be able to create and engage with it in ethically responsible ways. When combined with participatory co-creation initiatives, AI literacy can empower citizens to become active contributors to an inclusive and plural information environment rather than passive recipients of technological change. Citizens should be equipped to recognise and critically assess synthetic content, and to create and engage with it in ethically responsible ways. Finally, the governance of synthetic media ecosystems should be grounded in **multi-stakeholder collaboration**, bringing together citizens, technologists, policymakers, and ethicists in continuous dialogue and evaluation.

Ultimately, these findings underscore that democracy in the digital age is no longer defined solely by access to information but also by the ability to interpret, question, and co-create it. SOLARIS project therefore positions synthetic media not merely as a technological challenge to be regulated but as a social and cultural opportunity to reinvigorate public participation, strengthen civic agency, and promote a more inclusive and pluralistic digital democracy, aligned with the European Commission's vision of trustworthy and human-centric AI.

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. Jankovska**		

Deliverable	**Number**	6.2	**Title**	Regulatory innovations and policy options for generative AI and digital democracy GANs
Work package	**Number**	6	**Title**	Enhance capacities for digital citizenship and democracy
Task	**Number**	6.3	**Title**	Regulatory innovations and policy options for generative AI and digital democracy

Date of delivery	**Contractual**	M36	**Actual**	M36
Status	version 1.3		☒ Final version	
Nature	☒Report ☐Demonstrator ☐ Other ☐ORDP (Open Research Data Pilot)			
Dissemination level	☒ Public ☐ Confidential			

Authors (partners)	Federica Russo (UU), Tommaso Tonello (UU), Maria Dolores Sanchez Galera(UC3M), Andrew MacIntyre (UvA),			
Responsible author	**Name**	Yasaman Yousefi		
	Partner	DEX	**E-mail**	Yasaman.yousefi@dexai.eu

Summary (for dissemination)	D6.2 examines how generative AI and synthetic media affect democratic participation, public trust, and the integrity of the digital public sphere. Drawing on the findings of the SOLARIS project, and its three Use Cases, as well as the SOLARIS book, it analyses the risks posed by AI-generated content while also

	highlighting its potential for civic engagement and social good. The report proposes a harm-based, multi-layered governance approach and policy recommendations that combine legal clarity, institutional coordination, media and AI literacy, and citizen participation. Rather than relying on purely technical solutions.
Keywords	Generative Artificial Intelligence, Synthetic Media and Deepfakes, Digital Democracy, Media and AI Literacy, Democratic Resilience, Policy Recommendations

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
22/01/2026	V1.3	Federica Russo (UU)	Final editing of the text
15/01/2026	V1.2	Yasaman Yousefi (DEX)	Addressed the reviewers' comments
05/01/2026	V1.1	Federica Russo (UU)	Revision of whole document
20/12/2025	V1	Yasaman Yousefi (DEX) and Federica Russo (UU)	Harmonisation of contributions

Acknowledgement

Disclaimer

Contents

List of tables

Abbreviations and acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
AIGC	AI-Generated Content
ANT	Actor, Network Theory
API	Application Programming Interface
C2PA	Coalition for Content Provenance and Authenticity
CFR	Charter of Fundamental Rights of the European Union
CNN	Convolutional Neural Network
DARPA	Defense Advanced Research Projects Agency
DSA	Digital Services Act
EC	European Commission
ECHR	European Convention on Human Rights
ECSA	European Citizen Science Association
ELM	Extreme Learning Machine
EU	European Union
EPRS	European Parliamentary Research Service
ETS	Exponential Smoothing
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
HCI	Human, Computer Interaction
ICT	Information and Communication Technology
LCA	Latent Class Analysis
LLM	Large Language Model
ML	Machine Learning
NGO	Non-Governmental Organisation
PDTQ	Perceived Deepfake Trustworthiness Questionnaire
RNN	Recurrent Neural Network
SDG	Sustainable Development Goal
SSRN	Social Science Research Network
UC	Use Case
VLOP	Very Large Online Platform
WP	Work Package

1. Introduction

1.1. Context and Objectives

The emergence of **Generative Artificial Intelligence (GenAI)** and large language or diffusion models has revolutionised how information, images, and narratives are produced and circulated. These systems can synthesize realistic audio, visual content that is often indistinguishable from authentic media. Such capabilities hold transformative potential for creativity, education, and civic participation, but also pose serious risks to trust, accountability, and the integrity of democratic discourse, and are potential spreaders of fake news (Weiner and Norden, 2023). SOLARIS Deliverables 2.1 and 2.2 described these challenges thoroughly, especially from the perspective of democratic engagement and socio-political stability. AI-generated content (AIGC) has the potential to erode trust, unsettle public discourse, and reshape the informational conditions on which democratic agency depends. D2.1 mapped how AI-generated media can infiltrate elections, weaponize reputations, amplify geopolitical tensions, and generate climates of uncertainty that weaken collective deliberation. D2.2 deepened the analysis by showing that trust in AI-mediated content is formed within socio networks: users, platforms, norms, and technologies co-produce meaning. GenAI has the potential to destabilise the fragile infrastructures of civic confidence on which democratic life and socio-political stability rest.

Within this shifting landscape, SOLARIS investigated how AIGC poses threats and challenges to democracy, and how these technologies also have the potential to contribute to democratic engagement and participation. We conducted three Use Cases (UCs, which reframed deepfakes as phenomena shaped by networks of humans, platforms, institutions, and technologies rather than isolated artefacts. **UC1** revealed how technical quality, emotional engagement, and personal attitudes shape the detection and sharing of deepfakes, and resulted in a validated Perceived Trustworthiness Scale of deepfake audiovisual content. **UC2** assessed the challenges of AIGC and risks of infodemic, and the regulatory innovations needed to mitigate them, through conversations with journalists and experts. **UC3** demonstrated, through citizen science co-creation, that AIGC can be redirected toward democratic empowerment, producing value-based content that strengthens media literacy, inclusiveness, and digital citizenship. Accordingly, Work Package 6 (WP6) explores how AIGC can either erode or enrich civic life.

D6.2 represents the culmination of this inquiry. It builds on prior results, particularly **D6.1**, which developed methodologies for the co-creation of value-based AIGC, and integrates the feedback and findings from **Use Cases 2 and 3**. D6.2 extends these insights into the **policy and regulatory realm**, identifying how governance frameworks can foster transparency, accountability, and citizen empowerment in generative AI ecosystems.

Details are given in D5.2, which extensively discusses aims, methodologies, and results of the Use Cases. Briefly put, Use Case 2 (UC2) provided an applied exploration of the risks posed

by AI-generated content through the simulation of deepfake circulation in professional media environments. By engaging journalists and institutional actors within real newsroom settings, UC2 examined how synthetic media challenges existing verification workflows, journalistic ethics, and democratic resilience. The findings highlighted both procedural vulnerabilities and the critical role of human judgment in identifying and mitigating disinformation. These insights offer an empirical foundation for shaping policy instruments that strengthen institutional preparedness, promote responsible media policy, and reinforce trust mechanisms essential for transparent information ecosystems.

Use Case 3 (UC3) approached AI-generated content from a constructive perspective, exploring its potential to enhance digital citizenship through co-created value-driven synthetic media. By involving citizens in participatory workshops, UC3 investigated how generative AI can be ethically repurposed to support education, awareness, and democratic engagement. The exercise foregrounded public expectations for transparency, emotional authenticity, and accountability in AI-generated communication. The findings from this UC demonstrate how inclusive design and citizen participation can inform governance strategies. More details on the SOLARIS Use Cases can be found in Deliverable 5.2: *Use cases Co-Creative Evaluation.*

1.2. Structure of the Deliverable

This deliverable is structured to progressively move from the internal findings of the SOLARIS project to a broader policy-oriented analysis of generative AI, synthetic media, and democratic governance.

Section 2 provides a synthesis of the main findings of the SOLARIS book *Deepfakes, Democracy, and the Ethics of Synthetic Media*. It offers a structured overview of the book's eight chapters, distilling their core conceptual, empirical, and normative contributions. The purpose of this section is twofold: first, to consolidate the interdisciplinary knowledge produced within SOLARIS; and second, to establish a shared analytical foundation for the regulatory and policy discussions that follow. Rather than reproducing the book's arguments in full, Section 2 highlights the key insights most relevant to democratic risks, vulnerabilities, and governance challenges posed by synthetic media.

Section 3 situates these findings within the current European regulatory landscape. It assesses existing legal frameworks relevant to generative AI and synthetic media, including the AI Act, the Digital Services Act, and data protection law, identifying both strengths and persistent gaps. The section focuses, in particular, on unresolved ambiguities, enforcement challenges, and areas where existing regulation struggles to account for the socio-technical complexity of AI-generated content.

Section 4 deepens the analysis by examining technical, institutional, and governance challenges associated with mitigation strategies. It discusses the generative, detection arms race, the limits of purely technical solutions, and the need for coordinated, network-based

governance approaches. Drawing on SOLARIS use cases and stakeholder engagement, the section highlights the interdependence between technology developers, platforms, institutions, and citizens.

Section 5 translates the preceding analysis into policy recommendations. It advances a shift in regulatory paradigm from narrowly policing truth claims toward fostering democratic resilience through literacy and institutional coordination, and harm-aware governance. The section articulates concrete recommendations aimed at protecting democratic processes, empowering citizens, and strengthening institutional capacities, including targeted measures for vulnerable groups.

Finally, **Section 6** concludes the deliverable by reflecting on the broader implications of generative AI for democratic life. It summarises key insights and outlines future directions for policy, regulation, and research, emphasising the need to balance innovation with ethical safeguards and democratic resilience.

2. Synthesis of the Findings of the SOLARIS Book

In this section, we summarise the SOLARIS book, walking the reader through the chapters. The book is organised into 8 chapters:

i) The opening chapter reconstructs the foundations and technical developments that make deepfakes possible to generate and circulate;
ii) The second chapter then shows how deepfakes circulate on social media, stressing statistical opportunities and challenges to segmenting online groups with respect to the propagation of deepfakes, and reflects on the role of traditional media and citizens in fighting disinformation;
iii) Chapter 3 highlights how deepfakes are part of complex socio-technical systems, and we advance a 'network' approach to analyse them. In the same chapter, we next argue that semiotics, and especially visual semiotics, offers a valuable lens to understand the power of persuasion carried by deepfakes;
iv) Chapter 4 then reports on our empirical studies into the psychology of deepfakes: we test a psychometric scale for AI trustworthiness and identify several factors beyond quality of audiovisuals that explain trust in the contents;
v) In chapter 5, the book continues by highlighting the potential threats that deepfake generation and circulation may have on democratic processes, as well as their potential for delivering 'good' and constructive messages to the general public;
vi) Chapter 6 next addresses the classification problem of fake news and deepfakes from an ethical and sociological perspective, offering ethical and legal considerations regarding privacy, personal reputation, and liability related to the use of deepfakes in social media, and looking at mitigation strategies and their impact on citizens' rights;
vii) Chapter 7 examines the EU's regulatory response to deepfakes, focusing on the AI Act and its interaction with the DSA, GDPR, and liability regimes. It introduces the "Limited-Risk Paradox," showing how deepfakes with potentially severe democratic harms are primarily governed through transparency and downstream measures, leaving structural vulnerabilities in enforcement and prevention.
viii) Finally, Chapter 8 synthesises the book's findings and reframes deepfakes as a test case for democratic governance under technological uncertainty. It argues for democratic resilience through coordinated, layered responses that combine legal regulation, technical safeguards, institutional capacity, and civic empowerment rather than isolated or purely reactive solutions.

Chapter 1: Unmasking the Illusion: The Tech Behind Deepfakes

The opening chapter of the SOLARIS book traces the evolution of artificial intelligence from its conceptual birth in the mid-twentieth century to the generative architectures that underpin today's deepfakes. Beginning with Alan Turing's seminal question *"Can machines think?"* (Turing, 1950), and with John McCarthy's formal coining of the term "artificial intelligence"

in 1956 (McCarthy et al. 2006), the chapter situates deepfakes within a longer trajectory of computational ambitions: from symbolic rule-based systems, through machine learning, to the breakthroughs of deep neural networks.

The narrative highlights how the exponential growth of computational resources and data in the early 2000s enabled deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), with the decisive turning point arriving in 2014 with Ian Goodfellow's Generative Adversarial Networks (GANs) (Goodfellow et al. 2014): this innovation opened the door to synthetic content creation of unprecedented realism.

Alongside GANs, alternative approaches such as Variational Autoencoders (Kingma & Welling 2013), Transformers (Vaswani et al. 2017), and Diffusion Models (Ho et al. 2020; Ramesh et al. 2021; Rombach et al. 2022) further revolutionised generative AI, enabling multimodal synthesis and cinematic-level realism.

The second half of the chapter focuses on the **deepfake pipeline:** it details the techniques of face-swap, lip-sync, face reenactment, and voice cloning, explaining how each manipulates visual or auditory features to produce convincing synthetic outputs. The democratisation of these tools is traced from early open-source frameworks such as DeepFaceLab and FaceSwap, through the accessibility afforded by Google's Colab, to today's commercial platforms like FakeYou, DeepSwap, HeyGen, Runway Gen-4, and Google's Veo 3. These platforms have eliminated technical barriers, making synthetic media creation available to non-experts while embedding varying degrees of ethical safeguards. The toolkit now includes advanced image manipulation systems such as Google's Nano Banana, integrated into Gemini, which allow conversational editing of images with photorealistic consistency.

The chapter explains how synthetic media is created and, by unpacking this process, it highlights current features and future challenges of AIGC realism in terms of public trust, content authenticity, and democratic communication.

Chapter 2: The Spread of Deepfakes in Digital Networks

The second chapter of the SOLARIS book examines the spread of deepfakes across digital networks, situating them within the broader dynamics of social media virality, where engagement-driven algorithms, influencer amplification, and coordinated bot activity combine to accelerate the circulation of falsified content. The chapter demonstrates how emotional triggers such as fear, anger, or patriotism are systematically mobilised to enhance engagement and ensure that synthetic content reaches massive audiences quickly (Ali Adeeb & Mirhoseini, 2023; Pennycook & Rand, 2019).

The analysis then recalls the challenges posed by platform governance: Facebook, with its vast user base and weaker moderation tools, often allows deepfakes to circulate unchecked, while X has introduced contextual notes that help debunk falsehoods but cannot prevent their rapid

spread. The *continued influence effect* (Lewandowsky et al., 2012) is particularly relevant: once a manipulated image or video has been seen, its impression lingers even after correction, underscoring the urgency of early-warning systems and preventive interventions. The chapter also notes the discontinuation of monitoring tools such as CrowdTangle, which has limited the ability of journalists and independent researchers to track disinformation flows (Gotfredsen & Dowling, 2024).

In this chapter, techniques such as logistic regression, latent class analysis, factor analysis, clustering, and structural equation modelling are presented as essential tools for characterising groups vulnerable to deepfakes based on their sociodemographic, motivational, and cognitive factors. Additionally, the interaction of these factors with digital literacy and ideological predispositions is investigated (Bhatnagar & Ghose, 2004; Kang et al., 2020; Outwater et al., 2003; Verma, 2013; Yan et al., 2018; Shete et al., 2021). Nonetheless, the chapter also shows how current models remain limited by their reliance on Western datasets and short-term analyses, calling for broader, cross-cultural, and longitudinal studies.

The chapter subsequently examines institutional and media mitigation strategies, drawing in particular on evidence from the first activity of the SOLARIS project's Use Case 2, which analyses the role of newspapers' targeted interventions in reducing the circulation of harmful synthetic media. It analyses the contribution of traditional media to the identification, debunking, and contextualisation of deepfakes, while also addressing the professional challenges that AI-generated disinformation creates for journalists and media practitioners. The discussion further extends to the role of civil society in limiting disinformation-related risks. By identifying current shortcomings primarily in terms of insufficient AI literacy opportunities, the chapter argues for bottom-up educational approaches that enable citizens to actively shape their own literacy needs and objectives. Combining perspectives from journalistic practice, social media research, and statistical analysis, the chapter provides a consolidated account of deepfake propagation and response mechanisms, with the aim of informing both scholarly debate and policy development on strategies for mitigating this emerging threat after its occurrence.

Chapter 3: Semiotics of Synthetic Media

This chapter introduces a dual framework for interpreting synthetic media: by combining generative semiotics with Actor Network Theory (ANT) (Latour 2005), it addresses both the internal coherence of the image and the external socio-technical networks that grant it persuasive force. While synthetic media are part of a long genealogy of visual manipulation, contemporary generative AI introduces radical discontinuities in the **speed, scale, and social diffusion** of these practices, shifting the ability to manipulate from a smaller group of specialists to a larger population of *prosumers* (user-producers) through easily accessible GenAI tools. The democratisation of visual forgery heightens the political stakes, making it easier for deepfakes to undermine social trust. To analyse the internal textual organization, the

chapter advocates for the use of generative semiotics to focus on **Isotopies** (i.e., the recurrence of motifs, words, metaphors or narrative cues) produced through AI-generated media, where inconsistencies in terms of blurred contours, unnatural textures, and suspicious features of people represented in deepfakes (the "uncanny valley" effect) can unmask an image's artificial nature.

Furthermore, the **Anchorage** analysis stresses the critical role of verbal elements (captions, hashtags) in guiding interpretation and determining deepfakes' success in inheriting the trustworthiness of non-GenAI media. Nonetheless, four critical case studies (the Pope Balenciaga meme, Lola Flores' resurrection (Bassano, G., & Cerutti, M., 2024), the use of deepfakes in television journalism, and a viral Will Smith meme) highlight the shift of deepfakes towards **flawless credibility**, therefore overcoming previous graphic limitations of AI-generated media.

The final section integrates the generative semiotics, ANT dual framework into practical instruments for prevention and democratic resilience, including a **Taxonomy of Fakery** (distinguishing between falsification of the level of expression and falsification of content) and a **Reception Matrix** (mapping users' ability to identify and interpret markers of fabrication). The conclusion emphasises that the effects of synthetic media depend less on the tools themselves than on the networks of actors and on the competence of the audience. Therefore, effective prevention must prioritise **semiotic literacy, transparent labelling regimes, and context-aware pedagogy** to strengthen the capacity for unmasking deceptive content.

Chapter 4: The psychology of deception: Why we believe deepfakes?

This Chapter discusses how the advances in artificial intelligence have enabled the creation of highly realistic deepfakes, although, as anticipated in Chapter 3, their influence ultimately depends on humans' ability to perceive and interpret them. In this context, the chapter highlights widespread overestimation of digital citizens' detection abilities, with online users performing poorly at reliably distinguishing authentic from manipulated videos, often achieving results right at or slightly above chance levels (Diel et al., 2024).

This chapter then introduces the construct of **perceived trustworthiness of deepfakes**, defined as the extent to which individuals perceive a video as authentic. The development and validation of the multilingual **Perceived Deepfake Trustworthiness Questionnaire (PDTQ)** (Plohl et al., 2024) established a two, factor structure: **trustworthiness of content** (the plausibility and credibility of the presented information and its source) and **trustworthiness of presentation** (the perceived realism of delivery, including technical quality and the behaviour of the person in the video).

The chapter determines that perception of deepfakes is heavily influenced by **individual differences** spanning sociodemographic, motivational, and cognitive factors, in line with findings from Chapter 2. For instance, with respect to the selected sample, sociodemographic

factors showed that **older individuals** were more inclined to trust manipulated videos regarding both content and presentation (Plohl et al., 2024). Motivational variables, as indicated by the theory of motivated reasoning (Kunda, 1990), revealed that higher **political conservatism** and higher **trust in media** (a phenomenon known as misplaced trust) were positively associated with the perceived trustworthiness of content, but not of its presentation (Plohl et al., 2024). Regarding cognitive factors, higher **media literacy** and **reflective thinking** were protective factors against content trustworthiness, whereas higher **'bullshit receptivity'** was a strong risk factor. Notably, most individual differences were consistently associated with the perception of content trustworthiness, but only age (risk factor) and **deepfake, specific knowledge** (protective factor) were associated with the perceived trustworthiness of presentation.

The chapter shows how perceptions of trustworthiness are not passive judgments: they extend to shape **downstream psychological outcomes (**including shifts in attitudes and **behavioural intentions** (Ajzen, 1991; Ajzen & Fishbein, 1972). Trustworthiness of content was found to be the strongest predictor of viral **behavioural intentions** (e.g. sharing) (Plohl et al., 2024). Furthermore, experimental studies showed that the perceived trustworthiness of deepfake content consistently predicted attitude changes across polarised topics like climate change and immigration, regardless of objective video quality or political alignment (Plohl et al., 2025a). While presentation contributes to realism, the plausibility and coherence of the message (i.e. content) appear to be the more decisive factors in shaping attitudes.

If deepfakes are a technological challenge, belief in deepfakes is a psychological one: this chapter shows that protecting the public requires both technological detection tools and **psychological interventions** addressing the perceptual, cognitive, and motivational factors underlying media belief. The PDTQ's ability to quantify believability (Plohl et al., 2024) can directly inform scalable, evidence-based policy and platform responses, such as guiding automated moderation and developing deliberation prompts for users.

Chapter 5: Democracy Distorted: Deepfakes as Political Weapons

Chapter 5 argues that affordable generative AI has put "perfect-forgery" tools in the hands of political operatives, allowing deepfakes to be produced, uploaded (at zero cost), and algorithmically amplified before they can be effectively fact-checked. This chapter argues that synthetic media poses a **structural and recurrent threat** to democracies because it collapses the evidentiary foundations necessary for public discourse, electoral accountability, and social trust. The resulting harm is both material and **epistemic** (deliberately muddied channels of knowledge, eroding shared judgment) and **political** (weakening the norms for collective self, government).

Importantly, elections emerge once again as pivotal and highly vulnerable targets. While high, profile deepfakes (e.g. 2023 Slovakian elections) receive scrutiny and are often debunked;

however, the potentially more dangerous threat lies in **"microfakes",** synthetic content targeting low-profile, local political settings that often go undetected and unscrutinised (Ascott, 2020). The existing information environment, dominated by social media's **attention economy** and clustered echo chambers (Lewis & Marwick, 2017; Chun, 2024), is fertile ground for disinformation, accelerating the spread and acceptance of synthetic narratives that undermine trust in traditional authorities (Vaccari & Chadwick, 2020).

The chapter further discusses how AIGC contributes to worsening **infodemic** scenarios and to hindering **epistemic processes** in the online context, at the same time increasing participatory deterrence: the mere possibility that any authentic footage might be fabricated encourages dishonest actors to dismiss evidence, the so, called **"liar's dividend"** (Chesney & Citron, 2018), further contributing to public cynicism and epistemic harms. AIGC-related harms are disproportionately experienced by **women and minority groups**. In this context, non-consensual deepfake pornography, which has historically targeted women (Adjer et al., 2019), has represented a potent tool of **participatory deterrence**, strategically increasing the personal and reputational costs for women contemplating political careers (Kovalčíková & Weiser, 2021). This mechanism amplifies structural inequalities and compromises the democratic ideal of equal participation (Fricker, 2007).

Finally, the chapter stresses the relationship between the proliferation of deepfakes targets and the weakening of democratic values: i) **transparency** is eroded because visual proof becomes negotiable; ii) **equality** suffers insofar as well-resourced actors gain a comparative advantage in shaping narratives, leading to **credibility deflation** among those lacking digital and AI literacy; and iii) **accountability** is weakened when officials can evade scrutiny by simply claiming incriminating evidence is fake.

The chapter concludes that the systemic threat posed by deepfakes necessitates a multi-layered policy and educational strategy. To restore transparency, equality, and accountability, the chapter points at the importance of a three-pronged approach:

1. **Binding Regulatory Frameworks:** The **EU's Digital Services Act (DSA)** already mandates very large online platforms to assess and mitigate systemic risks from manipulated media. The **Artificial Intelligence Act (AI Act)** reinforces this by requiring mandatory, machine-readable **labelling** of synthetic images, audio, or video depicting real people, aiming to restore minimum transparency (Directorate-General for Communications Networks, Content and Technology, 2025). Furthermore, several member states have introduced criminalization and high penalties for non-consensual deepfakes and non-consensual labelling.
2. **Technological and Institutional Safeguards:** Efforts like Defense Advanced Research Projects Agency (DARPA)'s Semantic Forensics program (SemaFor) and

voluntary industry standards (e.g. Midjourney blocking prompts of political figures) raise the cost of deception.

3. **Educational Interventions:** Policy emphasises citizen-empowering **media literacy** through age, age-appropriate curricula, the OECD/EC's AI-literacy framework (Schleicher, 2025), and platforms like EUvsDisinfo, which foster source checking and critical reading.

Chapter 6: Synthetic Media for Social Good: Unlocking Positive Potential

This chapter investigates the ethical, communicative, and societal experiences of a series of participatory workshops (Use Case 3) conducted in collaboration with the European Citizen Science Association (ECSA). Engaging 44 participants from 18 national backgrounds, the workshops explored the potential of 'positive deepfakes' to promote social awareness of the UN Sustainable Development Goals (SDGs). UC3 aimed to understand how artificial texts are constructed and interpreted when used for educational and civic engagement purposes, rather than manipulation. Specifically, UC3 promoted citizen participation in co-creating AI content to advance **SDG 4 (Quality Education)** by fostering digital literacy and **SDG 16 (Peace, Justice, and Strong Institutions)** by promoting transparency and accountability in AI governance.

The co-creation process involved designing eight videos featuring: a synthetic version of Marie Curie, a synthetic character named Amina discussing the climate crisis, and a synthetic character named Casey advocating for mental health. These videos allowed for the development of two semiotic-informed taxonomies:

1. the **Textual Taxonomy** classified positive deepfakes based on their **discursive form** (i.e. the level of engagement, more detached or more intimate, through which the AI avatar conveys the message), **identity function** (using AI to either conceal and anonymise physical features of real people or to create hyper, realistic avatars, such as Marie Curie's posthumous deepfake), and **destination** (with AI being used to approach either a broader or a narrower, more targeted audience).
2. The **interpretive taxonomy** classified the different modes of audience reception, demonstrating that credibility does not depend solely on technical realism. Five primary modes emerged: i) **plastic interpretation** (technical realism, favoured by younger participants), ii) **discursive interpretation** (narrative coherence, favoured by older participants), iii) **ethical, cognitive interpretation** (appropriateness of use, e.g., acceptable in a museum but disturbing in a commercial context), iv) **passional interpretation** (emotional coherence of the media message), v) and **metareflective interpretation** (critical discussion of the technology's political implications). From these deliberations, four determinants of perceived impact emerged: native, intentional coherence (merging the first two points), technical, mimetic realism, ethical

transparency, and contextual adequacy, forming a basis for design and ethical guidelines.

The central policy implication of this Chapter is the need for a targeted approach to **Digital Education** that moves beyond the "deficit model", which simply trains citizens with the goal of increasing their technical skills. Instead, the findings support an **'AI education'** strategy (Panciroli & Rivoltella, 2023) focused on arousing **critical thinking** and fostering conscious, responsible digital citizenship. This approach emphasises that policy interventions must address both the technology and the interpretive capacity of the citizen at the same time. The Interpretive Taxonomy, which shows how different user groups prioritise different factors (content vs. technical quality), offers specific, evidence-based guidance for policymakers on creating effective, **transparent labelling regimes** and designing **context-aware pedagogical** tools. Ultimately, the chapter contends that the credibility of 'AI for Good' and its potential to promote social good and democratic communication depend on the transparent, intentional, and ethical framework guiding **how and why we choose to use it**. Policy must therefore ensure that diverse voices are included in the development of AI tools to prevent reinforcing systemic barriers to participation. In conclusion, while acknowledging the risks associated with deepfakes, it is possible to envision their ethical and responsible use. If properly regulated and oriented toward pro-social aims, deepfakes can become tools that support communication, education, and inclusion, thereby contributing positively to the promotion of human flourishing.

Chapter 7: Guarding the Truth: Assessing Current Mitigation Strategies

This chapter has three aims: i) it addresses the classification problem of fake news and deepfakes from an ethical and sociological perspective, framing synthetic media in terms of relational responsibility and the material consequences of immaterial harms; ii) it offers legal considerations regarding privacy, personal reputation, and liability related to deepfakes in social media; and iii) it examines mitigation strategies and their impact on citizens' rights, highlighting structural challenges in deepfakes governance.

The Chapter argues that online interactions between technology and users generate externalities either enhancing human capabilities and identity or eroding skills, autonomy, and personhood. Responsibility therefore cannot be reduced to individual legal liability: it must also encompass a dialogical capacity to relate to others and to the socio-technical ecosystem, a relational responsibility current overlooked by best-practices. The same rapid scale of information flows producing cognitive overload also undermines digital citizens' duty of care and justifies systemic risk obligations for Very Large Online Platforms (VLOPs) under the Digital Services Act (DSA), which redistribute relational responsibility between users and platforms. Yet, deepfakes' pervasiveness and misuse of personal data (see the case of Cambridge Analytica) intensify privacy and reputational harms. In this context, pornographic deepfakes are among the most notorious example of how synthetic media can inflict severe and lasting damage.

Next, the chapter stresses that mitigation requires constitutional balancing between fundamental rights (notably freedom of expression) and protections for democratic integrity, dignity, and pluralism. In the EU this balance must respect the European Convention on Human Rights (ECHR) and the Charter of Fundamental Rights (CFR): freedom of expression (Article 10 ECHR; Article 11 CFR) is not absolute and may be limited to protect dignity and pluralism (Article 10(2) ECHR). The European Court of Human Rights recognises heightened online risks, rendering EU regulatory responses constitutionally permissible and sometimes necessary. The chapter then highlights current gaps in complementary EU legislation tackling deepfakes. Article 17 of the General Data Protection Regulation (GDPR), dealing with the right to erasure, exemplifies a structural problem: deletion requests may not prevent AI models from retaining informational traces enabling re-synthesis of likenesses. Additionally, classification of many deepfakes as "limited, risk" under the AI Act, aimed at protecting innovation and legitimate expression, risks enabling uses incompatible with the democratic order and may shift verification burdens onto less AI-literate groups, weakening protections for those most vulnerable. Furthermore, the DSA imposes content, moderation, transparency, and systemic risk duties, particularly for VLOPs, but remains constrained: it is primarily reactive, insufficiently covers non-VLOP actors (e.g. private messaging), and depends on inconsistent VLOP cooperation.

Soft law instruments may play a vital role in providing timely responses and empirical evidence in a rapidly evolving domain, and it can inform and be integrated into binding frameworks to reduce semantic ambiguity, to translate principles into operational duties and, eventually, to tackle accountability gaps, stemming from legislative lag, algorithmic opacity, attribution difficulties, and semantic ambiguity about "accountability." Williams and others (Williams et al., 2022) propose operationalising accountability through five concepts (transparency, interpretability, explainability, reviewability, and traceability) mapped on two axes:

- The chronological axis, distinguishing between i) *ex ante* obligations (design and documentation requirements before deployment, linked to transparency and interpretability) and ii) *ex post* obligations (responses after events or behaviours, linked to explainability, reviewability, and traceability).

- The activity axis, demarcating between "push" obligations (information automatically provided) from "pull" obligations (information that must be actively requested).

This mapping yields concrete, enforceable duties. Soft-law instruments, such as the 2022 Strengthened Code of Conduct, can represent those push elements (e.g. periodic transparency reports on synthetic content, standardised reporting templates, electoral cooperation), but reporting from private stakeholders has often been inconsistent. Embedding such elements in binding frameworks like the DSA can provide legal context to those operational requirements introduced by Williams et al.'s framework. At the same time, this allows to translate ethical principles, such as relational responsibility, to legal obligations. At the same time, however,

EU-level "pull" dynamics remain underdeveloped: future AI literacy initiatives should extend beyond technical skills to foster contextual and content literacy, echoing SOLARIS Use Case 2 follow-up findings.

Despite a dense toolbox, deepfakes expose structural weaknesses that undermine democratic integrity and individual dignity. The chapter stresses the following five regulatory shortcomings as most critical:

i. Technological, Legislative Asymmetry: while regulation relies on detection and provenance (e.g. watermarking), adversaries can strip metadata, transcode files, re-edit outputs, or mount adversarial attacks.

ii. The Honest, Actor Problem and Transnational Enforcement Deficits: malicious actors often operate transnationally or via decentralised systems, evading EU jurisdiction.

iii. Fragmentation and Enforcement Deficit: Member States vary in enforcement capacity and interpret obligations divergently across data protection authorities, digital services coordinators, and courts.

iv. Insufficient Coverage of Immaterial Harms: erosion of dignity, emotional distress, and reputational harm remain inadequately addressed by EU law.

v. Uneven Impact and Distributive Vulnerability: technologically neutral rules may neglect distributive justice, disproportionately harming marginalised groups.

The Chapter concludes that constitutional traditions and case studies confirm that AI-amplified disinformation poses a systemic threat to democratic participation. Effective governance must be rooted in European pluralism and the right to informed choice, moving beyond a transparency-only paradigm. Priorities include public investment in interoperable detection and tamper-resistant provenance, international regulatory harmonisation, and expedited remedial mechanisms for reputational and psychological harm, which are the foundations for the policy recommendations in Chapter 8.

Chapter 8: Regulatory innovations and policy options for synthetic media and digital democracy

Drawing upon, and responding to, Chapters 5 and 7, the final chapter develops regulatory innovations for addressing deepfakes and generative AI in democratic contexts. The Chapter revolves around three main dimensions: i) legal instruments designed to address the malicious use of deepfakes; ii) a wider set of policy interventions targeting social harms linked to synthetic media; and iii) forward-looking strategies leveraging on generative technologies to strengthen democratic resilience. The chapter proposes a framework reconciling technological innovation with security and democratic safeguards, and it offers best-practice

recommendations for policymakers, technology developers, and media actors. It also acknowledges ethical, institutional, and practical constraints such as infrastructural shortcomings, governance failures, tensions with freedom of expression, and normative questions about permissible uses of synthetic media.

The chapter starts from the premise that contemporary societies exist in an infosphere where human experience and knowledge are reconstituted through information flows (Floridi, 2014). AIGC thus alters beliefs, as well as the very structural conditions under which societies construct, verify, and contest knowledge (Russo, 2022; Bisconti et al., 2024; McIntyre et al., 2025). The 2024 US presidential elections, characterised by a widespread presence of AIGC, is representative of such societal harms through erosion of institutional trust. Current frameworks, such as the EU AI Act, acknowledge that GenAI can produce material or immaterial harms, but prevailing responses remain largely reactive, emphasising moderation and detection over systemic interventions.

Emerging policy strategies cluster into three categories:

i) Retreat strategies aimed at reducing digital interactions in favour of in-person exchange to rebuild trust: these are impractical at scale and risk forfeiting the positive democratic affordances of digital technologies.
ii) Containment strategies focused on detecting, labelling, and limiting harmful AIGC: many such approaches adopt an explicit "policing truth" orientation, regulatory, technical, or discursive measures that attempt to suppress misinformation, but this emphasis risks eroding pluralist debate and can have authoritarian tendencies.
iii) Mobilisation strategies that leverage on GenAI itself to strengthen democratic processes.

Drawing on Feinberg's account of harm as wrongful obstruction of interests (Feinberg, 1987) and Smuha's work (2021), the Chapter breaks down harms associated with AIGC, categorising them on three levels: (i) individual, when people are directly misled (e.g. deceptive deepfakes); (ii) collective, when groups are disproportionately affected (e.g. racial stereotypes); and (iii) societal, when institutions and governance are undermined (e.g. synthetic media in elections). From a complementary perspective, political theorists such as Mouffe, Laclau, and Rancière suggest the pathologies often diagnosed as "post-truth" reflect deficits in meaningful democratic participation rather than mere scarcity of "facts". From these inputs, the Chapter claims that counter-disinformation should not be limited to fact-policing: it must be coupled with interventions that expand political participation; thus, "more politics" rather than "more truth". Given the diffusion of media, production capabilities to ordinary citizens, policy responses must combine AIGC containment with citizens' mobilisation: technical measures and institutional coordination to limit harm, and proactive deployment of generative tools to enhance representation, deliberation, and civic engagement.

Chapter 8 emphasises AIGC instances that concretely improve public engagement or empathetic communication. Examples include translating government communications for multilingual communities (e.g. Manoj Tiwari speaking Haryanvi in 2020), interactive exhibits that explain history (e.g. Dalí Lives), visualisations clarifying abstract policy consequences (e.g. This Climate Does Not Exist), and EXHIBIT A, i (Blackburn 2023), which used GenAI to visualise witness statements of 32 refugees held on Manus Island and Nauru. Importantly, in EXHIBIT A, the synthetic images were explicitly acknowledged as artificial and not presented as evidence, an instructive model for ethical representational uses. With broader access to GenAI, citizens can more easily document and communicate lived experiences (including systemic violence or neglect) in forms that resonate with wider publics, thereby amplifying marginalized voices.

The Chapter advocates for how democratic resilience strategy should address harms across Smuha's three levels and be organised around three key objectives: i) legal clarification of AIGC and informational harms; ii) coordination of democratic institutions; and iii) promotion of a plural and participatory citizenship. These objectives foreground both defensive measures (protecting rights and institutions) and proactive investments (civic adoption of generative tools).

To reduce ambiguity and enable consistent enforcement, Chapter 8 argues that the EU should harmonise relevant rules across jurisdictions and update core instruments: firstly, the EU Copyright Directive should be updated to give citizens a right to one's own likeness similarly to performers. Secondly, the GDPR should be updated to redefine AIGC that replicates an individual's likeness or voice as protected personal data, even if created entirely synthetically. Finally, the AI Act's transparency obligations could be expanded to include individual consent and rapid takedown rights. Together, these updates would strengthen protections against misrepresentation through AIGC and close regulatory gaps.

Because harms occur across borders, regulatory harmonisation must be complemented by capacity, building in lower-resourced regions. The digital divide at both the citizen and national levels, low digital literacy, limited infrastructure, and weaker institutions, undermines any unified approach. To tackle the divide, responses should include: (i) authentication systems (digital watermarks, provenance registries) that enhance verifiability and evidentiary uses in courts; coordinated investment in detection and prevention AI tools; (ii) diplomatic measures and international agreements to deter state actors, reinforced where appropriate by economic sanctions; and (iii) investments in foundational digital infrastructure and regulatory capacity through international cooperation and assistance programmes.

Harmful AIGC spreads rapidly through networks, so early-warning systems integrating technical detection and human intelligence are essential. To bridge data, sharing hindrances between the private and the public sectors, policy should enable neutral third, party convenors and a shared set of ethical principles for collaboration. A unified counter-disinformation

network, linking government institutions, social media platforms, fact-checkers, and media organisations at multiple levels could support a real-time infodemic alert mechanism whereby identified AIGC is flagged for collective review and simultaneous public awareness campaigns.

Central to resilience is strengthening local journalism and trusted community media. Increased funding for local outlets enables reliable, firsthand reporting that feeds higher-level responses and serves as a trusted intermediary. Policies might establish national or international funds, combining government grants and philanthropic contributions, to supply smaller media organisations and civil society groups with advanced tools and training to counter capacity gaps.

AI literacy should be integrated into formal curricula to teach citizens how to detect AIGC (e.g. unnatural eye movements, distorted backgrounds, audio glitches), and to analyse production processes, underlying biases, and contextual provenance. Literacy programmes should emphasise contextual evaluation (source, background information) rather than sole reliance on technical artefacts. Education should also promote ethical, pro-democratic uses of AIGC, practical skills in AI-enabled personal representation and civic expression to cultivate a population that is both critical and participatory.

These recommendations must navigate constraints: enforcement capacity varies across states; detection tools produce false positives; counter-disinformation and transparency requirements must navigate their relationship with the right to free expression; and restraint is needed to avoid measures that centralise epistemic authority. Policymakers must balance targeted legal clarification and coordinated institutional responses with robust protections for pluralist debate. Where trade-offs are unavoidable, priority should be given to interventions that are proportionate, right-respecting, and that reinforce rather than displace democratic contestation.

In short, a sustainable policy approach treats AIGC neither as an exclusively technical problem to be contained nor solely as a civic deficit to be corrected by truth-policing. Instead, it brings together legal clarification and institutional coordination with investments in media, education, and participatory uses of generative technologies. This integrated strategy, defensive where necessary and mobilising where possible, seeks to mitigate harms at the individual, collective, and societal levels while leveraging AIGC's communicative affordances to strengthen pluralist democratic practice.

3. Existing Regulatory Frameworks: An Assessment of Legal Gaps and Strengths

The next sections further expand on the topics addressed throughout the SOLARIS book, and especially chapter 8. Importantly, they further elaborate on and develop effective policy recommendations. To address necessary regulatory steps, a recap of the existing legislative framework to tackle GenAI risks at the EU level is presented first.

3.1. The EU's Legislative Response to Generative AI

The European Union has an ambitious and evolving legislative framework to address digital and information-based harms. While no single piece of legislation was exclusively designed for deepfakes, several key regulations collectively form a response.

- **The AI Act (AIA):** AIA is the world's first comprehensive legal framework for AI, adopting a risk-based model in which stricter obligations apply to systems classified as high-risk. Article 50(4) introduces mandatory transparency obligations for deepfakes, requiring providers and deployers to disclose that content has been artificially generated or manipulated. However, under a genuinely risk-based approach, the regulatory burden should correspond to the magnitude of the risks posed. Yet deepfakes, despite the Act's own acknowledgement of their capacity to manipulate behaviour and undermine democratic processes (Recitals 28, 29, 133), were ultimately placed in the limited, risk tier rather than the high, risk or prohibited categories. This creates what scholars describe as a limited risk paradox: although deepfakes pose systemic dangers to elections, public trust, and democratic discourse, the Act responds primarily with transparency duties rather than substantive safeguards or strong oversight. As a result, the AIA's reliance on reclassification mechanisms and provenance disclosure is difficult to interpret and enforce, and insufficiently calibrated to the democratic harms the technology can produce. The Act therefore leaves a significant gap: it lacks clear enforcement pathways or meaningful remedies for victims, and its limited, risk designation undercuts the regulatory ambition needed to address deepfakes' destabilising societal impact.
- **The Digital Services Act (DSA):** This regulation aims to create a safer online environment by implementing measures to combat illegal content and harmful activities. It places a strong emphasis on transparency and accountability for very large online platforms, requiring them to mitigate systemic risks related to disinformation. A significant limitation, however, is its reactive "notice and action" mechanism. This means that platforms are only required to act *after* the harmful content has been distributed and the damage has been done. Additionally, the DSA has a limited scope, as it explicitly excludes private messaging services like WhatsApp, or other social media platforms, which are a major avenue for deepfake distribution.

- **The General Data Protection Regulation (GDPR):** The GDPR's foundational principles of data protection and the right to rectification provide a basis for challenging deepfakes created using an individual's personal data, such as their face or voice. However, the "household exemption" leaves a significant portion of the population unprotected, as the regulation does not apply to personal data processed for purely personal or household activities.

- **Product Liability Directive (PLD).** It is also important to note that PLD has been revised in parallel with these regulatory processes, introducing measures designed to mitigate the information asymmetry between producers and users of AI systems. The revised Directive treats AI software as a "product" and introduces disclosure, burden, shifting, and transparency obligations (Articles 9, 13), helping victims establish liability in cases of AI-related damage (Novelli et al., 2024). The scope of the Directive has been extended to include all AI systems and AI-enabled goods (excluding open-source software unless integrated into commercial products), reflecting the EU's recognition of AI's opacity and the imbalance of information between developers and consumers. This step represents an important breakthrough in adapting liability rules to the realities of generative AI and large language models. However, PLD reveals marked limitations when applied to deepfakes. While it reduces evidentiary burdens for victims and acknowledges AI models as legally relevant products, its remedial focus remains oriented toward physical injury and property damage. Non-material harms, such as reputational injury, dignity violations, or psychological distress, remain undercompensated. This means that although the GDPR offers direct pathways to challenge unlawful deepfake processing, PLD provides only partial remedies and relies heavily on the AI Act to fill liability gaps (Novelli et al., 2024; Hacker et al., 2023).

Despite these shortcomings, the EU's regulatory stack also exhibits important strengths that form a solid foundation for future governance. The AI Act's introduction of a legal definition of deepfakes and its embedding of provenance and disclosure rules represent an important first step toward harmonised European standards. The DSA, for its part, institutionalises systemic risk assessments and establishes clear governance mechanisms, including Digital Services Coordinators and enhanced oversight for VLOPs that create unprecedented pathways for monitoring, auditing, and enforcing platform duties. Meanwhile, the GDPR provides a powerful rights-based framework for contesting unlawful data use, anchoring deepfake harms in existing privacy and dignity protections and empowering supervisory authorities to intervene proactively. These instruments demonstrate the EU's capacity to build an integrated, multilayered governance architecture capable of evolving in parallel with technological developments. They show that the regulatory infrastructure for future, more dynamic safeguards is already in place; what remains is strengthening their coordination, closing substantive gaps, and ensuring that governance mechanisms operate coherently across the AI lifecycle.

3.2. Identified Regulatory Deficiencies and Unresolved Ambiguities

An analysis of the current legal landscape reveals a fundamental mismatch between the nature of the harm and the nature of the regulatory response. Existing laws are fragmented, reactive, and often misaligned with the systemic, epistemic threats posed by GenAI.

The EU's legislative approach, particularly the AI Act, relies heavily on a transparency obligation and reactive content moderation. The AI Act's disclosure requirement is described as relatively lenient given the potential for harm. This is an insufficient remedy, as it does not minimize the harm to a victim, especially in cases of non-consensual deepfake pornography where the damage is immediate, irreversible, and not mitigated by post-hoc labelling. Likewise, the DSA's reactive mechanism means the harm has already occurred before any action can be taken, making it a system for damage control rather than prevention. These limitations reflect deeper structural causes: the absence of proactive detection mechanisms, the lack of dynamic and continuously updated media, literacy, and dissemination initiatives, and a regulatory framework that still treats synthetic media primarily as a content-moderation issue rather than a systemic risk to democratic participation. As a result, the current legal regime largely manages the outputs of the problem rather than addressing the upstream drivers that allow harmful deepfakes to proliferate at scale (Fisher et al., 2024).

A major unresolved issue is the legal uncertainty created by fragmented national laws and the open-to-interpretation wording of the AI Act. With countries like Spain, and Germany taking different national approaches to complement initiatives from Brussels, this creates a "patchwork of regulations" that malicious actors can easily exploit by operating from the least regulated environments. In EU legal terms, such fragmentation also generates significant "spillover effects": regulatory gaps in one Member State can weaken enforcement across the entire Union, allowing harmful content, services, or actors to circulate freely through the internal market despite stronger protections elsewhere. These spillovers dilute national safeguards and even undermine the harmonising purpose of EU law itself, making coordinated action indispensable. The lack of a unified legal definition for deepfakes and its associated harm creates a significant loophole, undermining the efforts of more proactive jurisdictions. A fragmented legal response to a global, rapidly evolving technology like deepfakes can lead to a race to the bottom in terms of regulation, creating a global vulnerability that no single nation can solve alone. The EU governance model is an attempt to foster innovation within structured environments, where experimentation and compliance evolve in tandem, but rule of law requirements are challenged by the dynamic of racing to expand and innovate that is intrinsic to the market mechanisms.

4. The Limits of Technical Solutions and the Prospects of the SOLARIS Network Approach

4.1. The Generative-Detection Arms Race

The allure of a purely technical solution to the deepfake problem is strong, but it is ultimately an ex-post strategy, perpetually lagging behind the occurrence of AI-related harm. The relationship between deepfake generation and detection is often described as an "arms race" (George and George, 2023). As generative models become more sophisticated, detection systems must infer manipulation from increasingly subtle artefacts in the output, often relying on datasets built from known falsifications.

Yet detection research has an inherent paradox: the more a detection model improves, the more it exposes the precise weaknesses that generative models must eliminate. In doing so, detection tools unintentionally create a feedback mechanism that accelerates the evolution of the fakes themselves. This adversarial dynamic ensures that generative capabilities consistently outpace defensive ones, particularly as open-source architectures make rapid iteration easier for malicious actors than for regulators or legitimate researchers. This is why detection alone cannot serve as a reliable cornerstone of an effective governance framework. It is technologically fragile and dependent on training data that become obsolete as soon as new manipulation techniques emerge. It is also structurally limited, as it places the burden of verification on platforms, journalists, or even end users.

Moreover, detection cannot address the broader epistemic consequences of synthetic media, such as the “liar’s dividend”, the erosion of evidentiary trust, and the systemic uncertainty that deepfakes introduce into democratic discourse. A comprehensive regulatory response, therefore, requires shifting from reactive technical solutions to upstream, structural measures: robust provenance and watermarking standards, interoperable audit mechanisms, proactive detection integrated earlier in the upload pipeline, and governance models that anticipate rather than merely respond to technological evolution. Without such proactive measures, a solely technical approach will remain locked in perpetual catch-up, creating an unstable and insufficient basis for safeguarding democratic processes and individual rights.

4.2. A Critical Examination of Technical Mitigation Measures

In the literature, two major technical measures for the mitigation of deepfake spreading have been identified: watermarking and the ‘Safe by Design’ principle.

- **Watermarking and Digital Provenance:** One of the most frequently cited technical responses to synthetic media is the use of invisible watermarks or provenance metadata standards, such as those developed by the Coalition for Content Provenance and Authenticity (C2PA). These approaches aim to embed information about origin and modification history directly into AI-generated content. However, **as standalone**

safeguards, such solutions face significant technical and practical limitations. Watermarks can be removed or degraded with relative ease, particularly in the context of open-source models, and may be lost through routine editing processes. Provenance metadata systems such as C2PA, while more robust at the infrastructural level, remain vulnerable to circumvention, for instance, through simple actions such as screen capturing, which strips associated metadata. Moreover, security research has demonstrated that motivated actors can generate falsified provenance records, raising the risk of a more pernicious failure mode in which inauthentic content is presented as verifiably genuine (Romanishyn et al, 2025).

- **The "Safe by Design" Principle in Practice:** The concept of "safe by design" aims to minimize the potential for misuse of generative AI to harm children, for example, by training generative models to proactively prevent the creation of Child Sexual Abuse Material. While tech companies are committed to these principles, effective prevention measures may vary depending on whether the models are open-source or closed-source, highlighting the complexity of implementation.

4.3. SOLARIS network approach to governing deepfakes

A core critique of a purely technical approach is that it risks falling into the fallacy of *technological solutionism*, whereby technology is framed simultaneously as the source of and the remedy for a fundamentally social and political problem. Mandating a technical measure that is not yet sufficiently mature or robust, such as supposedly "difficult to remove" watermarks, can create an **illusion of safety without delivering its substance**. If watermarking is presented as a decisive guarantee of authenticity, the public may develop a false sense of security, inferring that content lacking a watermark is necessarily genuine. In such cases, the information environment may become more fragile rather than more resilient, as critical scepticism is replaced by misplaced trust in a flawed or partial technical signal.

Furthermore, relying on technical fixes to solve a political and social problem can displace political responsibility. The EU AI Act's primary mechanism for deepfake regulation is a transparency obligation, which is a technical and informational solution. This approach, however, reinforces the idea that democracy is dependent on private tech companies and their systems, shifting the burden of defining and addressing harm from public discourse into the hands of algorithms and corporate policy. This fundamental philosophical conflict suggests that current EU policy, while comprehensive in scope, may be built on the flawed premise that the problem is a technical one when it is, in fact, a much broader one, in which safeguarding democratic processes requires a political and social response as well.

SOLARIS is critical of the idea that deepfakes can be meaningfully governed by focusing on the artefact *alone*, whether through detection tools, watermarking schemes, or transparency labels. This is because these measures overlook the "socio-technical systems" in which synthetic media is produced, circulated, interpreted, and weaponised. Building on the Actor-

Network Theory (ANT) framework adopted in D2.2, SOLARIS conceptualises deepfakes not as nodes within a dynamic web of human actors (users, creators, victims, regulators), technical artefacts (models, platforms, recommendation systems), and institutional structures (media ecosystems, legal frameworks, political cultures) (Bisconti et al, 2024; McIntyre et al, 2025). In this configuration, harm arises from the interactions, incentives, and asymmetries within the wider network that enable manipulation, accelerate dissemination, and shape user trust. This approach shows that any governance model that targets the technical artefact alone will inevitably fall short, because it regulates only the visible symptom and not the relational system that produces and amplifies it. At the same time, SOLARIS shows that deepfakes can be used ethically in political, educational, and health-related contexts when embedded in appropriate institutional and governance frameworks, underscoring that their impact depends on context rather than the artefact itself.

The Network approach we argue for shows that governing deepfakes requires intervening at different joints of the network: upstream at the point of model design and training; midstream at the level of platform architecture, recommender dynamics, and content framing; and downstream at the level of public literacy, institutional resilience, and democratic norms. We emphasise that trust in AIGC is co-constructed through these interdependencies and not determined merely by the technical quality of the media. Thus, governance must include systemic measures like strengthened platform responsibilities, transparent provenance ecosystems, participatory regulatory design, and continuous, socially embedded digital literacy initiatives. SOLARIS therefore reframes the deepfake challenge as fundamentally political and relational, requiring democratic, institutional, and socio-cultural responses in addition to technical safeguards. In doing so, it provides a future-ready governance model that recognises deepfakes as socio-technical assemblages rather than technical anomalies and treats democracy not as something to be protected by tools alone, but as a networked practice requiring continuous collective stewardship.

5. Policy Recommendations: Advancing Pro-Democratic AI Governance

This section showcases what SOLARIS has learned and translate that into actionable policy recommendations.

SOLARIS engaged in three use cases to understand the psychological mechanisms of trust in deepfakes (UC1), the role of media in deepfakes detection and debunking (UC2), and the potential of AI-generated content for delivering 'good' messages (UC3). The Use Cases are thoroughly described and evaluated in D5.2. Collectively, they highlight a dual policy imperative: to **protect democratic ecosystems** from synthetic manipulation by reinforcing professional verification, cross-sector coordination, and regulatory clarity; *and* to **empower citizens** through participatory governance, literacy, and ethical co-creation of AI content. This dual mandate reflects the balance between institutional protection and civic empowerment that underpins resilient democracies in the age of generative AI.

A robust policy response to deepfakes must move beyond a reactive, legalistic framework and adopt a proactive, multi-layered approach that addresses systemic harms, in accordance with the network approach developed by SOLARIS. This requires a philosophical shift from the "post-truth" discourse, which seeks to re-establish a single source of truth, to one that cultivates a vibrant, pluralistic, and genuinely political public sphere (Farkas and Shou, 2023). As Farkas and Schou argue, democracy thrives not through consensus but through "agonistic pluralism," where divergent perspectives coexist and contend within shared democratic norms. This shift moves the focus from containment to cultivation, from controlling the informational environment to empowering citizens within it.

This conceptual reframing was exemplified in UC3, which demonstrated that generative AI could serve as a civic instrument rather than a threat. By involving citizens in assessing synthetic media, UC3 turned participants into active interpreters and ethical agents, transforming them from passive consumers of information into co-authors of meaning. Participants identified that transparency, contextual framing, and emotional authenticity were key to fostering trust and engagement. This approach reflects the "more politics" model (Farkas & Schou, 2023): rather than seeking to eradicate misinformation through censorship or algorithmic policing, it empowers citizens to interpret, challenge, and co-create, turning generative AI into a participatory medium for democratic learning.

5.1. A Shift in Paradigm: From Policing 'Truth' to Cultivating Literacy, Politics, and Humanities

Drawing on data from the European Parliamentary Research Service (EPRS, 2021) and on preceding analyses, our proposed policy framework advances three overarching objectives:

(i) **Clarification**: harms and legal responsibilities need to be better clarified than currently done in the available legal framework.
(ii) **Coordination**: institutional coordination and infrastructural capacity need to be further strengthened, with a harmonised European approach; and
(iii) **Citizenship:** citizenship through literacy, participation, and pluralism needs to be enhanced through dedicated programmes**.**

Each of these objectives is articulated through specific policy recommendations designed to address Smuha's three levels of harm (individual, collective, and societal) while cultivating a more inclusive and participatory information ecosystem. We build here on the approach proposed by Smuha (Smuha, 2021), who distinguishes harm at three levels:

- **Individual Harm:** This occurs when a deepfake directly deceives or misleads a single person, impairing their ability to make informed decisions. The abovementioned case of financial fraud case in Hong Kong is an example where an individual's agency was directly compromised by a convincing impersonation (Milmo, 2024). For victims of non-consensual deepfakes, the harm is also deeply personal, causing severe psychological trauma, anxiety, and even post-traumatic stress disorder.
- **Collective Harm:** This emerges when deepfakes disproportionately affect specific groups, reinforcing and amplifying existing biases and inequalities. AI-driven image generators, for example, have been shown to amplify racial biases by yielding images dominated by white individuals when prompted for certain professions. This reflects existing societal prejudices and risks entrenching them further by normalizing skewed representations. Additionally, the effectiveness of deepfake detection tools is often compromised by these biases, with error rates for individuals with darker skin tones being higher.
- **Societal Harm:** This form of harm extends beyond individual and collective grievances to impact the structural integrity of public institutions and democratic governance itself. The proliferation of deepfakes, as seen during the US election cycle, can deepen public uncertainty, erode trust in democratic processes, and undermine the rule of law. This systemic pollution of the infosphere makes it more difficult for society to construct, verify, and contest knowledge, thereby weakening the foundations of democratic deliberation and shared reality.

In the following sections, we comment and further expand on each of the three principles we identified.

Table 1 shows how the three principles previously identified (Clarification, Coordination, Citizenship) help address the three levels of risk (individual, collective, societal).

Table 1 Priority proposals for democratic resilience

Objectives	Priority proposals	Harm level		
		Individual	Collective	Societal
Clarification	**Unified legal framework**	**(x)**	**(x)**	**(x)**
	Unified personality rights	(x)		
	Transparency obligations	(x)		
Coordination	Unified infrastructural investment			(x)
	Multi-stakeholder coordination			(x)
Citizenship	Media and AI literacy	(x)	(x)	
	Technical citizenship	(x)	(x)	(x)
	Pluralist media landscape		(x)	(x)

Policymakers could draw on these insights by supporting initiatives that reflect similar principles and operational logics, including:

- **Structured professional dialogues and simulations**: For example, regular cross-sector tabletop exercises involving journalists, platform trust and safety teams, electoral authorities, and civil society organisations could simulate the emergence of a high-impact political deepfake during an election period. Such initiatives can strengthen shared interpretive standards, clarify responsibilities, and improve cross-sector preparedness without relying exclusively on automated detection.
- **Participatory co-creation and evaluation programmes**: These programmes can be embedded in public education, cultural institutions, or citizen science initiatives, enabling citizens to develop interpretive and ethical capacities alongside technical understanding.
- **Evidence-informed literacy frameworks**: Policymakers could integrate such frameworks into existing digital education strategies, professional training schemes, and public awareness campaigns.
- **Harm-aware governance mechanisms**: This includes tailored responses for victims of non-consensual deepfakes, safeguards against group-based misrepresentation, and systemic protections for electoral and informational integrity.

Taken together, these approaches reflect a shift from reactive containment toward **democratic cultivation**: strengthening the capacity of institutions and citizens alike to navigate, interpret, and contest synthetic media within a pluralistic public sphere. In this sense, the SOLARIS experience offers a set of **transferable governance principles** that can inform concrete policy initiatives aimed at balancing innovation, participation, and democratic resilience in the age of generative AI.

5.2. Protecting Vulnerable Groups through a Harm-Level Approach

Equipped with a harm-level approach, regulatory and policy interventions can more effectively identify, prevent, and mitigate the disproportionate impacts of generative AI and synthetic media on vulnerable groups. The SOLARIS project has systematically analysed how AI-generated content produces differentiated harms across individual, collective, and societal levels. Rather than assuming uniform exposure or impact, SOLARIS demonstrates that harms are unevenly distributed and often reinforce existing structural inequalities.

In brief, the groups identified as particularly vulnerable include the following:

- **Women and girls** are disproportionately targeted by non-consensual deepfake pornography and sexualised synthetic content, leading to reputational damage, psychological harm, and participatory deterrence in public and political life.
- **Political candidates, journalists, and public figures**, especially at local or marginal levels, are exposed to reputational manipulation through "microfakes" that evade scrutiny and fact-checking.
- **Elderly populations and individuals with lower levels of digital and media literacy** are more susceptible to believing and sharing deceptive synthetic content.
- **Marginalised and minority communities**, who are more likely to be targeted through stereotypical, discriminatory, or identity-based synthetic narratives, are amplifying collective and group-level harms.
- **Citizens in highly polarised or crisis contexts**, where emotional manipulation through synthetic media exacerbates distrust, infodemic dynamics, and institutional erosion.

The harm-level approach enables these vulnerabilities to be addressed in a structured and proportionate manner. At the **individual level**, it supports targeted protections such as expedited takedown mechanisms, clear consent and personality rights safeguards, access to redress, and psychosocial support for victims of synthetic media abuse. At the **collective level**, it justifies tailored interventions for groups facing systemic exposure, including enhanced monitoring of discriminatory content, group-sensitive risk assessments under the Digital Services Act, and specialised media literacy programmes designed for specific demographic and cultural contexts. At the **societal level**, the approach underpins regulatory obligations aimed at safeguarding democratic processes and institutional trust, including transparency and provenance requirements, election-period safeguards, and coordinated responses to large-scale disinformation campaigns.

By explicitly linking regulatory responses to differentiated harm levels, this framework avoids one-size-fits-all solutions and instead promotes **proportionate, distributive, and dignity-centred governance**. In doing so, it aligns with the EU's fundamental rights framework and the SOLARIS network approach, ensuring that the protection of vulnerable groups becomes a

core criterion in the design, deployment, and oversight of generative AI systems rather than a secondary or reactive concern.

5.3. Strengthening Institutional Coordination and Infrastructural Capacity

A comprehensive and future-oriented response to deepfakes and generative manipulation requires a whole-of-society approach grounded in inter-institutional collaboration and multi-level governance. UC2 revealed that fragmented infrastructures and siloed responsibilities weaken the collective capacity to identify and respond to synthetic media threats. Journalists, regulators, and platforms often operate independently, relying on ad hoc communication rather than systematic coordination. To address this gap, the EU should facilitate the creation of a real-time, multi-stakeholder alert and verification network linking media organizations, fact-checkers, regulatory agencies, and civic actors. When a deepfake or other form of generative manipulation is detected by one, whether audiovisual, textual, or multimodal, it should trigger a coordinated review across all participating institutions. Crucially, such a system would not require that every deepfake trigger an automatic cross-institutional response; rather, it would provide stakeholders with a shared escalation infrastructure that can be activated when the content is assessed to pose significant public interest, democratic, or individual rights risks. In this way, the mechanism enables timely, proportionate, and concerted action when deemed necessary, transforming today's fragmented and reactive responses into a proactive, networked defence. This approach aligns with the Digital Services Act's systemic risk assessment framework and the European Media Freedom Act's emphasis on newsroom resilience and coordinated threat response.

This "whole-of-society" coordination must also be accompanied by **targeted infrastructural investment**. This should be complemented by capacity-building programmes that address Europe's **digital asymmetries**, ensuring that smaller or under-resourced Member States can access the same verification and authentication capabilities as larger ones. In doing so, the EU would protect the integrity of democratic communication and strengthen **epistemic equity** across the Union.

At the core of these efforts lies the **empowerment of citizens through literacy**. UC1 confirmed the importance of literacy in the psychometric scale. Similarly, UC2 and UC3 both emphasized that the ability to recognise and critically interpret synthetic content extends far beyond detecting visual or auditory anomalies. **AI and media literacy** must integrate **argumentation and critical reasoning skills**, enabling individuals to assess *how* content is made, and *why* it was created, *who* benefits from it, and most importantly, *what contextual cues* shape its credibility. The follow-up study for UC2 highlights the importance of literacy about contextual factors, such as political narratives, cultural frames, and institutional trust that influence how synthetic information is received and acted upon.

We can therefore assert that plausibility, not technical perfection, is what makes deepfakes effective. For journalists and experts, the most reliable first step in identifying synthetic media is therefore not pixel-level analysis but attention to the context in which the content appears: which online channels disseminate it, whether it aligns with existing political narratives, who stands to gain from its circulation, and how it fits within broader cultural and informational frames.

AI and media literacy must therefore integrate argumentation and critical reasoning skills. Literacy about contextual, narrative, and institutional factors is as important as understanding the technical mechanisms behind deepfakes. Strengthening these forms of literacy equips both citizens and professionals to interpret synthetic content as part of a communicative ecosystem whose meaning and impact are co-produced by its context.

Therefore, literacy should be understood as **a multimodal competence**, encompassing audiovisual, textual, and discursive forms of AI-generated manipulation. Citizens must learn to analyse persuasive strategies, rhetorical framing, and narrative coherence, developing an awareness of the socio-political intentions embedded within generative content. This also includes understanding algorithmic curation and the role of platform infrastructures in amplifying or containing such material. Democratic resilience depends not merely on recognising falsehoods but on cultivating **contextual judgment and argumentative maturity,** which are skills that enable citizens to navigate the infosphere as reflective participants rather than passive recipients.

5.4. Protect Democracy: Legal and Legislative Recommendations

This section sets out a set of legal and legislative recommendations aimed at safeguarding democratic processes from the harms associated with synthetic media, while ensuring coherence across the existing EU regulatory landscape. Building on the harm-based framework developed earlier in this deliverable, the section is structured around three complementary pillars: (a) the need for a unified legal framework capable of addressing synthetic media consistently across jurisdictions and policy domains; (b) the harmonization of personality and identity rights to protect individuals from non-consensual synthetic reproduction; and (c) transparency obligations designed to support accountability and institutional oversight without creating misplaced trust in technical solutions. Together, these measures seek to strengthen legal certainty, protect fundamental rights, and preserve epistemic trust in democratic communication.

5.4.1. Unified Legal Framework on Synthetic Media

Journalists and institutional actors operate in an environment of uncertainty when confronting synthetic media. They rely on context and professional judgment rather than clear legal standards or technological guarantees. This underscores the need for a unified EU legal framework that clearly defines AIGC and distinguishes it from traditional disinformation. A

taxonomy that differentiates between falsification of form (e.g., manipulated video) and falsification of content (e.g., misrepresentation of authentic material) would enhance legal clarity and enforcement consistency. Such a framework should align with the AI Act, GDPR, and national counter-disinformation laws to ensure coherent governance of synthetic media across Member States.

In UC2, journalists expressed difficulty in attributing responsibility when manipulated media spread across borders and platforms. Clarifying liability is therefore crucial for creators and intermediaries such as platforms, AI model providers, and political campaigners who deploy synthetic media in electoral contexts. **The AI Act's provisions on transparency and the Digital Services Act's (DSA) due diligence obligations should be expanded to include explicit duties for detecting and labelling AIGC.** A unified framework would thus address Smuha's three levels of harm, protecting individuals from reputational injury, collective institutions from disinformation, and society from the erosion of epistemic trust.

a) Unified Personality Rights

UC3 raised complex ethical questions regarding likeness, consent, and posthumous representation. Participants expressed discomfort with deepfakes that resurrect deceased figures or mimic living individuals, even when used for noble ends such as public education or awareness campaigns. Several participants reported feeling somewhat uneasy about the deepfakes shown in UC3, both those portraying a revived public figure and those involving a synthetic "ordinary person." Even when intended for education or awareness, these materials were seen as problematic in terms of accountability. To address these forms of harm, the EU should harmonize personality rights, covering likeness, voice, and image, by recognizing identity as a protected form of intellectual property. Drawing inspiration from the Danish Copyright Act amendment entered into force in the Summer of 2025, EU legislation could extend copyright, style protections to all citizens, enabling legal recourse against unauthorized synthetic reproductions.

Specifically, **the Copyright Directive could be amended to protect all likenesses**, not just those of performers; the GDPR should classify synthetic reproductions of identifiable persons as personal data, even when derived from generative models rather than recordings; and the **AI Act should require explicit consent for likeness** use and ensure rapid takedown procedures. These measures would safeguard individuals (from non-consensual manipulation), collectives (such as cultural or political groups misrepresented by AI), and society as a whole (by reinforcing trust in communicative authenticity).

b) Transparency Obligations

Both UC2 and UC3 converged on transparency as the cornerstone of trustworthy AI governance. UC2 participants noted that journalists identified deepfakes primarily through contextual reasoning rather than automated detection, **calling for transparent labelling systems that allow immediate recognition of synthetic content.** Similarly, UC3 participants

emphasized that disclaimers and provenance information increased acceptance and trust, provided they were persistent and contextually clear.

Policy should therefore require the use of visible labels and provenance-related metadata for AI, AI-generated or AI-altered material, establishing these measures as **baseline transparency obligations** rather than as guarantees of authenticity. Such disclosures should indicate, where feasible, the involvement of AI systems, the nature of the generation or alteration, and relevant provenance information, thereby operationalising the transparency requirements set out in Article 52 of the AI Act.

These mechanisms should primarily serve **institutional functions**, supporting regulatory oversight, platform accountability, journalistic verification, and research, rather than placing the burden of authentication on individual users. In this regard, public or semi-public registries of AI, AI-generated content, administered by competent EU or national digital authorities, could complement labelling requirements by enabling authorised actors to trace content origins, assess compliance, and investigate harmful uses.

Transparency, understood in this way, is a procedural safeguard that underpins accountability and informed scrutiny. At the same time, and as discussed above, transparency measures alone are insufficient to address the risks associated with deepfake production and circulation. They must therefore be embedded within a broader governance framework that includes harm-based assessments, platform responsibilities, institutional coordination, and sustained investment in media and AI literacy.

5.5. Empower Citizens: Enhancing Citizenship through Literacy, Participation, and Pluralism

This section is structured around three mutually reinforcing pillars for strengthening democratic resilience in the age of generative AI. We begin by addressing **Media and AI Literacy**, outlining how expanded literacy frameworks can empower citizens to critically interpret AI-generated content without relying on technical verification tools. We then introduce **Technical Citizenship** as a participatory model that positions citizens as active, ethical contributors to AI-mediated public life , enabling them with defensive strategies through literacy. Finally, the need for a **Pluralist Media Landscape** is examined, detailing how targeted public support, certification, and inclusive AI use can sustain media diversity, editorial independence, and democratic participation in an increasingly synthetic information environment.

a) Media and AI Literacy

Empowering citizens through education emerged as a central finding across both UC2 and UC3. UC2 showed that professional expertise and contextual judgment allow journalists to resist deception even in the absence of advanced detection tools, highlighting the importance

of cognitive resilience rather than purely technical safeguards. UC3 extended this insight to the public sphere, demonstrating that citizens who understand the mechanisms, intentions, and aesthetic conventions of generative AI are better equipped to interpret synthetic media critically and to engage with it responsibly.

Policy should therefore expand traditional media literacy to explicitly include **AI literacy,** as a defence mechanism. This entails enabling citizens to recognise, contextualise, and critically assess AI-generated content (AIGC) by understanding how meaning, credibility, and persuasion are constructed. Such literacy must encompass interpretative reasoning, ethical reflection, and awareness of the socio-technical environments in which content circulates. The European Commission's Digital Education Action Plan already provides an appropriate framework for integrating these competencies across formal education, professional training, and lifelong learning pathways.

To safeguard pluralism and avoid the risk of state monopolisation of truth, these initiatives should be led and implemented by **independent institutions**, including universities, NGOs, public interest organisations, and citizen science networks. Literacy, in this sense, functions as a defensive mechanism against deception and even a **democratic skill set**: a foundational capacity for informed participation in the digital public sphere.

Citizens do not require professional-grade verification tools. What they benefit from instead is a simplified and actionable interpretative framework: understanding which platforms and channels are more prone to the circulation of manipulated content; recognising when a media item aligns too seamlessly with polarising or emotionally charged narratives; and knowing when to pause, seek corroboration, or withhold judgment. Effective literacy programmes should therefore integrate basic source, checking practices, ethical awareness, and knowledge of platform dynamics alongside technical understanding of AI.

b) Techno-Ethical Citizenship

Beyond literacy, democratic resilience increasingly depends on what can be described as **technical citizenship**. This concept reflects a broader shift in democratic expectations: citizens demand protection from harmful AI uses, as well as meaningful opportunities to shape how AI technologies are developed and deployed within society. Technical citizenship thus repositions individuals not merely as consumers or risk bearers, but as **ethical creators, communicators, and participants** in AI, AI-mediated public life.

Publicly funded programmes should support this shift by enabling civic engagement with generative AI through initiatives such as civic hackathons, open-source creative laboratories, community-based AI workshops, and participatory design environments. These spaces allow citizens to experiment with generative tools in transparent and responsible ways, fostering both technical familiarity and ethical awareness.

When embedded within clear ethical guidance and democratic oversight, generative AI can expand citizens' expressive capacities. It can help articulate lived experiences, support empathetic storytelling, enhance accessibility, and enable more inclusive participation in political and civic dialogue. This approach aligns with the AI Act's emphasis on human oversight and with Horizon Europe's commitment to co-creation and citizen science. By cultivating technical citizenship, policy frameworks shift from a logic of control to one of **participatory resilience**, strengthening democracy through engagement rather than restriction.

c) Pluralist Media Landscape

A resilient democratic ecosystem ultimately depends on a **diverse, independent, and pluralist media environment**. As synthetic media accelerates the pace and complexity of information production, news organisations face mounting pressures to verify content, maintain editorial standards, and preserve public trust. Supporting smaller, local, and independent media actors is therefore essential, as these organisations play a critical role in sustaining epistemic diversity, local accountability, and representational balance.

Policymakers should promote participatory initiatives such as citizen science projects, literacy programmes, and Living Labs by providing targeted grants, specialised training, and equitable access to verification, provenance, and transparency tools. Such measures strengthen the media ecosystem's collective capacity to adapt to technological fluidity, while ensuring that no single institution, public or private, becomes the central arbiter of truth. The objective is not centralisation, but a **distributed and resilient information environment** capable of resisting manipulation and disinformation.

Within this framework, public funding can also incentivise **certified organisations** to deploy AI-driven content production in ways that amplify underrepresented voices and linguistic minorities. These organisations would include public service media, non-profit and community media outlets, accredited journalism organisations, civil society actors, cultural and educational institutions, and public interest digital platforms whose mandates explicitly encompass democratic participation, inclusion, or fundamental rights protection.

Certification should not function as a content, approval, or truth validation mechanism. Rather, it should operate as a **process-based and value-based accreditation**, grounded in transparent criteria. These would include: (i) compliance with relevant EU legal frameworks, notably the AI Act, the Digital Services Act, and the GDPR; (ii) the adoption of internal governance standards for the ethical use of generative AI, including transparency, traceability, and meaningful human oversight; (iii) demonstrable editorial independence and accountability structures; and (iv) a clear commitment to pluralism, non-discrimination, and public interest objectives. Certification could be administered by independent public authorities or delegated bodies such as national media regulators, data protection authorities, or EU-level coordination

mechanisms, building on existing oversight practices rather than creating new bureaucratic layers.

Linking public funding to such certification would serve two complementary purposes. First, it would lower structural barriers that currently prevent smaller, non-commercial, or minority language organisations from accessing and responsibly using generative AI tools. Second, it would actively encourage the development of AI-generated content oriented toward inclusion, such as multilingual public information, local democratic reporting, accessibility, enhancing formats, and participatory or co-created media initiatives. In this way, generative AI is positioned not as a force of concentration or homogenisation, but as a **means of redistributing communicative power**.

A pluralist media landscape supported along these lines reduces structural asymmetries of attention and enhances representational justice. It addresses harms at the individual, collective, and societal levels by mitigating exclusion, countering narrative monopolies, and strengthening citizens' exposure to diverse perspectives. The future of democratic communication thus depends both on constraining harmful uses of generative AI and on embedding these technologies within an ecosystem of **diversity, transparency, and civic agency**, where public-interest actors are structurally enabled to shape the digital public sphere.

6. Conclusions: Balancing Innovation, Ethics, and Democratic Resilience

Generative AI confronts democracies with an ethical paradox. It multiplies opportunities for expression and inclusion while simultaneously amplifying the vectors of deception. The challenge is not to prohibit synthetic media but to embed them within a normative architecture that sustains democratic values. The SOLARIS project's contribution lies in demonstrating that such embedding is possible when technological innovation proceeds hand in hand with civic participation and regulatory foresight. The methodologies tested in Use Case 3 revealed how citizens can meaningfully shape AI outputs, ensuring that generative systems reflect public values rather than market incentives.

The policy path forward requires integrating regulation, education, and design. European policymakers should reinforce transparency and accountability within the AI Act's framework, promote public AI literacy initiatives, and institutionalize co-regulation through participatory oversight bodies. Civil society and academia must continue to mediate between citizens and technology, fostering inclusive debate and ethical reflection. Platforms and developers bear responsibility to implement provenance tracking, user-empowerment features, and rigorous auditing of generative models. Together, these measures constitute the foundation of a resilient digital democracy.

Ultimately, the question posed by SOLARIS is not whether democracy can survive synthetic media, but whether it can evolve through them. By treating synthetic actors as partners in, rather than adversaries to, the public sphere, we may reclaim technology as a medium of solidarity and shared meaning. In this sense, regulatory innovation is inseparable from cultural imagination: both are required to turn the age of generative AI into an era of democratic renewal.

References

Adjer, H., Giorgio, P., Francesco, C., & Laurence, C. (2019). *The state of deepfakes: Landscape, threats, and impacts.*

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179, 211. https://doi.org/10.1016/0749, 5978(91)90020, T

Ajzen, I., & Fishbein, M. (1972). Attitudes and normative beliefs as factors influencing behavioural intentions. *Journal of Personality and Social Psychology, 21*(1), 1, 9. https://doi.org/10.1037/h0031930

Ali Adeeb, R., & Mirhoseini, M. (2023). The impact of effect on the perception of fake news on social media: A systematic review. *Social Sciences, 12*(12), 674.

Bassano, G., & Cerutti, M. (2024). Posthumous Digital Face: A Semiotic and Legal Semiotic Perspective. *International Journal for the Semiotics of Law, Revue internationale de Sémiotique juridique, 37*(3), 769, 791.

Ben Doherty (2023) 'Manus Island and Nauru: Previously Unseen Testimony and AI Imagery Reveal "Unimaginable" Part of Australian History'. *The Guardian* 8 April 2023. Retrieved from: https://www.theguardian.com/australia, news/2023/apr/08/manus, island, and, nauru, previously, unseen, testimony, and, ai, imagery, reveal, unimaginable, part, of, australian, history.

Bhatnagar, A., & Ghose, S. (2004). A latent class segmentation analysis of e-shoppers. *Journal of Business Research, 57*(7), 758, 767.

Bisconti, P., McIntyre, A., & Russo, F. (2024). Synthetic socio-technical systems: Poiêsis as meaning making. *Philosophy & Technology*, *37*(3), 94.

Brigitte Tousignant (2021) This Climate Does Not Exist: Picturing impacts of the climate crisis with AI, one address at a time. *Mila*. 13 October 2021. Retrieved from https://mila.quebec/en/article/this, climate, does, not, exist, picturing, impacts, of, the, climate, crisis, with, ai, one, address, at

Chesney, R., & Citron, D. (2019, February). *Deepfakes and the new disinformation war: The coming age of post-truth geopolitics.* Foreign Affairs. https://www.foreignaffairs.com/articles/world/2018, 12, 11/deepfakes, and, new, disinformation, war

Chesney, R., & Citron, D. K. (2018). Deep fakes: A looming challenge for privacy, democracy, and national security. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3213954

Chun, W. H. K. (2024). Discriminating data: Correlation, neighborhoods, and the new politics of recognition. MIT Press.

de Ceglia F.P., (2018) Aldilà digitale. Come internet ha cambiato la nostra percezione della morte.

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta, analysis of 56 papers. *Computers in Human Behavior Reports, 16*, 100538. https://doi.org/10.1016/j.chbr.2024.100538

EPRS (European Parliamentary Research Service). (2021). *Tackling deepfakes in European policy*. Retrieved from:

https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

European Commission. (2023). *European Declaration on Digital Rights and Principles for the Digital Decade* (2023/C 23/01). Official Journal of the European Union, C 23/1.

Farkas, J., & Schou, J. (2023). Post-truth, fake news and democracy: Mapping the politics of falsehood. Routledge.

Feinberg, J. (1987). *Harm to others*. Oxford University Press.

Fisher, S. A., Howard, J. W., & Kira, B. (2024). Moderating synthetic content: The challenge of generative AI. *Philosophy & Technology*, *37*(4), 133.

Floridi, L. (2010). The Information Society and its Philosophy. *The Information Society*, *26*(3), 153, 158.

Floridi, L. (2014). The Fourth Revolution: How the Infosphere is Reshaping Human Reality. Oxford University Press.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

George, A. S., & George, A. H. (2023). Deepfakes: the evolution of hyper realistic media manipulation. *Partners Universal Innovative Research Publication*, *1*(2), 58, 74

Gotfredsen, S. G., & Dowling, K. (2024). *Meta is getting rid of CrowdTangle, and its replacement isn't as transparent or accessible. Columbia Journalism Review.* Retrieved October 11, 2025, from https://www.cjr.org/tow_center/meta, is, getting, rid, of, crowdtangle.php

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1112, 1123).

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. Advances, *Neural Information Processing Systems*, 33, 6840, 6851. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b, Abstract.html

Kang, L., Ye, S., Jing, K., Fan, Y., Chen, Q., Zhang, N., & Zhang, B. (2020). A segmented logistic regression approach to evaluating change in caesarean section rate with reform of birth planning policy in two regions in China from 2012 to 2016. *Risk Management and Healthcare Policy, 13*, 245, 253. https://doi.org/10.2147/RMHP.S230923

Kingma, D. P., & Welling, M. (2013). *Auto, Encoding Variational Bayes*. https://openreview.net/forum?id=33X9fd2, 9FyZd

Kovalčíková, N., & Weiser, M. (2021, August 30). Targeting Baerbock: Gendered disinformation in Germany's 2021 federal election. *Alliance for Securing Democracy.* https://securingdemocracy.gmfus.org/targeting, baerbock, gendered, disinformation, in, germanys, 2021, federal, election/

Kultur Ministeriet (2025) Broad agreement on deepfakes gives everyone the right to their own body and their own voice. 26 June 2025. Retrieved from https://kum.dk/aktuelt/nyheder/bred, aftale, om, deepfakes, giver, alle, ret, til, egen, krop, og, egen, stemme

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480, 498. https://doi.org/10.1037/0033, 2909.108.3.480

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford University Press.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106, 131. https://doi.org/10.1177/1529100612451018

Lewis, B., & Marwick, A. E. (2017). *Media manipulation and disinformation online.* Data & Society Research Institute. https://datasociety.net/library/media, manipulation, and, disinfo, online/

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazin*e, 27(4), 12, 12.

McIntyre, A., Conover, L., & Russo, F. (2025). A network approach to public trust in generative AI. *Philosophy & Technology*. *(Advance online publication, update when vol./issue available)*

Milmo, D. (2024, February 5). *Company worker in Hong Kong pays out £20m in deepfake video call scam*. The Guardian. https://www.theguardian.com/world/2024/feb/05/hong, kong, company, deepfake, video, conference, call, scam

Mouffe, C. (2013). Agonistics: Thinking the World Politically. Verso.

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law & Security Review*, *55*, 106066.

Novelli, Claudio, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. 2024. 'Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity'. *SSRN Electronic Journal*, ahead of print. https://doi.org/10.2139/ssrn.4694565.

O'Halloran, A. (2021). The Technical, Legal, and Ethical Landscape of Deepfake Pornography. Cs. Brown. Edu.

Outwater, M. L., Castleberry, S., Shiftan, Y., Ben, Akiva, M., Zhou, Y. S., & Kuppam, A. (2003). Attitudinal market segmentation approach to mode choice and ridership forecasting: Structural equation modeling. *Transportation Research Record, 1854*(1), 32, 42. https://doi.org/10.3141/1854, 04

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39, 50.

Panciroli, C., & Rivoltella, P. C. (2023). Pedagogia algoritmica: Per una riflessione educativa sull'intelligenza artificiale.Scholé.

Plohl, N., Mlakar, I., Aquilino, L., Bisconti, P., & Smrke, U. (2024). Development and validation of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ) in three languages. *International Journal of Human, Computer Interaction, 41*(11), 6786, 6803. https://doi.org/10.1080/10447318.2024.2384821

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero, Shot Text, to, Image Generation, *arXiv*. https://doi.org/10.48550/arXiv.2102.12092

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. 10674, 10685.

Romanishyn, A., Malytska, O., & Goncharuk, V. (2025). AI-driven disinformation: Policy recommendations for democratic resilience. Frontiers in Artificial Intelligence, 8, Article 1569115. https://doi.org/10.3389/frai.2025.1569115

Rousay, V. (2023). Sexual deepfakes and image-based sexual abuse: Victim, survivor experiences and embodied harms (master's thesis, Harvard University).

Russo, F. (2022). Techno, scientific practices: An informational approach. Rowman & Littlefield.

Schleicher, A. (2025, April 29). New AI literacy framework to equip youth in an age of AI. *OECD Education and Skills Today.* https://oecdedutoday.com/new, ai-literacy, framework, to, equip, youth, in, an, age, of, ai/

Shete, A., Soni, H., Sajnani, Z., & Shete, A. (2021). Fake news detection using natural language processing and logistic regression. *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, 136, 140. https://ieeexplore.ieee.org/abstract/document/9563292/

Smuha, N. A. (2021). Beyond the individual: Governing AI's societal harm. *Internet Policy Review, 10*(3). https://doi.org/10.14763/2021.3.1579

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *49*, 433–460.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1), 205630512090340. https://doi.org/10.1177/2056305120903408

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa, Abstract.html

Verma, J. P. (2013). Cluster analysis: For segmenting the population. In *Data analysis in management with SPSS software* (pp. 317, 358). Springer India. https://doi.org/10.1007/978, 81, 322, 0786, 3_10

Weikmann, Teresa, and Sophie Lecheler. 2024. 'Cutting through the Hype: Understanding the Implications of Deepfakes for the Fact-Checking Actor, Network'. Digital Journalism 12 (10): 1505, 22. https://doi.org/10.1080/21670811.2023.2194665.

Weiner, M. D. (2023). Destined to Deceive: The Need to Regulate Deepfakes with a Foreseeable Harm Standard. *Michigan Law Review*, *122*(4).

Williams, R., Cloete, R., Cobbe, J., Cottrill, C., Edwards, P., Markovic, M., ... & Pang, W. (2022). From transparency to accountability of intelligent systems: Moving beyond aspirations. *Data & Policy*, *4*, e7.

Yan, S., Kwan, Y. H., Tan, C. S., Thumboo, J., & Low, L. L. (2018). A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Medical Research Methodology, 18*(1), 121. https://doi.org/10.1186/s12874, 018, 0584, 9