Deepfakes, Democracy, and the Ethics of Synthetic Media

A Synthesis of the SOLARIS Project

Edited by: Yasaman Yousefi, Lucy Conover,
Izidor Mlakar, and Federica Russo

University of Maribor Press

University of Maribor

Faculty of Electrical Engineering
and Computer Science

Deepfakes, Democracy, and the Ethics of Synthetic Media

A Synthesis of the SOLARIS Project

Editors

Yasaman Yousefi

Lucy Conover

Izidor Mlakar

Federica Russo

February 2026

Title Deepfakes, Democracy, and the Ethics of Synthetic Media

Subtitle A Synthesis of the SOLARIS Project

Editors Yasaman Yousefi
(DEXAI, University of Bologna, CIRSFID ALMA AI)

Lucy Conover
(Utrecht University, Freudenthal Institute)

Izidor Mlakar
(University of Maribor, Faculty of Electrical Engineering and Computer Science)

Federica Russo
(Utrecht University, Freudenthal Institute)

Review Antonio Carnevale
(University of Bari “Aldo Moro”)

Mariya Yurukova
(University “St. Kliment Ohridski”)

Fabiana Ciccarella
(University of Bari “Aldo Moro”)

Language editing Shelagh Hedges

Technical editors Marina Bajić
(University of Maribor, University of Maribor Press)

Jan Perša
(University of Maribor, University of Maribor Press)

Cover designer Lucy Conover
(Utrecht University, Freudenthal Institute)

Graphic material Sources are own unless otherwise noted.
The authors and Yousefi, Conover, Mlakar, Russo (editors), 2026

Published by University of Maribor
Založnik University of Maribor Press
Slomškov trg 15, 2000 Maribor, Slovenia
https://press.um.si, zalozba@um.si

Issued by University of Maribor
Izdajatelj Faculty of Electrical Engineering and Computer Science
Koroška cesta 46, 2000 Maribor
https://www.feri.um.si, feri@um.si

Edition	First edition
Publication type	E-book
Available at	http://press.um.si/index.php/ump/catalog/book/1089
Published at	Maribor, Slovenia, February 2026

CIP - Kataložni zapis o publikaciji
Univerzitetna knjižnica Maribor

004.946:321.7(082)(0.034.2)

DEEPFAKES, democracy, and the ethics of synthetic media [Elektronski vir] : a synthesis of the SOLARIS project / editors Yasaman Yousefi ... [et al.]. - 1st ed. - E-zbornik. - Maribor : University of Maribor, University of Maribor Press, 2026

Način dostopa (URL): https://press.um.si/index.php/ump/catalog/book/1089

ISBN 978-961-299-109-8 (PDF)
doi: 10.18690/um.feri.2.2026
COBISS.SI-ID 267279363

ISBN 978-961-299-109-8 (pdf)
978-961-299-110-4 (hardcover)

DOI https://doi.org/10.18690/um.feri.2.2026

Price Free copy

For publisher Prof. Dr. Zdravko Kačič,
Rector of University of Maribor

Attribution Yousefi, Y., Conover, L., Mlakar, I., Russo, F. (eds.). (2026). *Deepfakes, Democracy, and the Ethics of Synthetic Media: A Synthesis of the SOLARIS Project.* University of Maribor, University of Maribor Press. doi: 10.18690/um.feri.2.2026

Table of Contents

Acknowledgements

This book was prepared as part of the activities of the SOLARIS project, funded under the European Union's Horizon Programme (Grant Agreement No. 101094665).[1] The views and opinions expressed in this book are those of the authors alone and do not necessarily reflect the official position of the European Union. Neither the European Union nor the granting authority can be held responsible for any use that may be made of the information contained herein.

The editors and authors gratefully acknowledge Dr Antonio Carnevale, Dr Mariya Yurukova, and Fabiana Ciccarella for their careful review of the manuscript and for their insightful comments, which significantly enhanced its scientific quality. We are also deeply indebted to the members of the SOLARIS Ethics Advisory Board**,** Dr Virginia Ghiara, Prof. Sigrid Sterckx, and Prof. Antonella Marchetti**,** for their invaluable guidance throughout the project, and in particular for their contributions to the design and development of the Use Cases.

SOLARIS organized numerous outreach and dissemination events, and we sincerely thank all participants for their questions, feedback, and engagement, which played a crucial role in shaping our research and refining our analyses.

[1] More details about the SOLARIS project can be found at https://projects.illc.uva.nl/solaris/

Last but not least, we are grateful to the University of Maribor Press for their professional assistance throughout the editorial, production, and dissemination stages of this publication.

University of Maribor Press

Introduction: The Rise of Deepfakes and the Need for an Interdisciplinary Approach

YASAMAN YOUSEFI,[1, 2] LUCY CONOVER,[3] FEDERICA RUSSO[3]
[1] DEXAI-Artificial Ethics, Rome, Italy
yasaman.yousefi@dexai.eu
[2] University of Bologna, CIRSFID ALMA AI, Faculty of Legal Studies, Bologna, Italy
y.yousefi@unibo.it
[3] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
l.a.conover@uu.nl, f.russo@uu.nl

In the digital age, disinformation has evolved from a peripheral, yet not new, issue into a structural and persistent threat. Driven by algorithmic amplification and advances in generative AI technologies, false and misleading content now permeates public discourse with unprecedented speed, outreach, and likeliness. At the centre of this transformation is synthetic media: content such as images, videos, audio, or text that is generated or manipulated by AI. This category includes deepfake videos, AI-generated voices, and hyper-realistic digital avatars, all of which blur the boundary between authenticity and fabrication, challenging traditional notions of evidence and trust in digital communication.

As artificial intelligence (AI) generated content becomes increasingly indistinguishable from authentic media, the question is no longer just whether we can tell the difference between AI-generated and real content, but what difference it makes if we cannot. This book invites scholars, policymakers, journalists, technologists, and citizens to grapple with that question, and to imagine a future where synthetic media supports, rather than subverts, democracy.

In a hyperconnected digital environment, the distinction between objective truth and subjective belief becomes increasingly blurred. Individuals are expected to verify content themselves, yet the volume, speed, and complexity of information make this task nearly impossible. This vulnerability becomes exploited by highly realistic, emotionally compelling content that often aligns with prior beliefs, making it difficult to distinguish reality from fabrication.

Understanding the threat of deepfakes requires looking at their impact on multiple levels. At the individual level, deepfakes can violate privacy and identity, causing reputational damage and psychological harm. At the collective level, they can intensify social polarization, target marginalized groups, and erode trust in institutions. At the societal level, deepfakes can destabilize public discourse, distort political decision-making, and weaken democratic processes. Addressing these risks requires more than technical solutions. It demands an integrated approach that includes legal frameworks, cultural literacy, and systemic safeguards to preserve trust and democratic integrity.

We are witnessing an increasing tension between synthetic authenticity and democratic integrity. Synthetic media, grounded in artificial intelligence, has already been used to fabricate political speeches, simulate attacks, and manipulate public figures, often blurring the line between satire and subversion. One image of Pope Francis in a Balenciaga coat was harmless and humorous to some; another deepfake video of President Donald Trump endorsing climate action, though well-intentioned, demonstrates how such simulations can dangerously alter perceived political positions. While technology offers novel possibilities for creativity, education, and inclusion, they also threaten the foundations of democratic deliberation by undermining trust in what is seen and heard. This duality, between creative empowerment and informational disorder, defines the stakes of this work and is at the heart of the book. The challenge is not only technical but also social and political, reflecting the ways in which synthetic content interacts with human perception, cultural norms, and institutional structures.

The volume synthesizes the research conducted in the European Horizon SOLARIS project, which began in February 2023 with a clear mandate: to understand the impact of Generative Adversarial Networks (GANs) and of other generative AI technologies able to manipulate and generate audiovisual content on democratic processes. Not only the threats, but also the opportunities. SOLARIS could take advantage of an international, interdisciplinary, and intersectoral consortium, involving organizations from seven European countries (plus the UK), including humanities, social sciences, and computer sciences, professional media and journalism, governmental organizations, and citizen science associations.

SOLARIS developed a theoretical framework to study the phenomenon of deepfakes, namely the "network approach". According to this approach, AI-generated content – the artefact – is a node in the broader socio-technical systems it is part of. It is, in fact, insufficient to merely consider AI-generated audiovisual contents qua artefacts, and for their technical qualities. The approach builds on Actor-Network Theory (a major approach in Science and Technology Studies), and is complemented with considerations from philosophy and ethics of technology, visual semiotics, and law. All these perspectives are needed in order to account for the full cycle, from design and development of the technologies making the generation of synthetic media possible, to the reception and shaping of public discourse. Any attempt to regulate and govern synthetic media in the hope of contrasting the growing phenomenon of disinformation must take all this complexity into account. We have conducted empirical research, developing a psychometric scale to measure "trust" in AI-generated content, we have engaged with professional journalists and with citizens to thoroughly analyse the potential dangers and opportunities of AI-generated media, and on this basis, we have assessed the existing regulatory framework.

This manuscript is a collection of eight interconnected chapters, each touching upon this topic from a different dimension, to provide a comprehensive understanding of the risks and opportunities of AI-generated content, combining technical insight, human psychology, and policy analysis. The central argument of the book is that the deepfake crisis is ultimately a crisis of trust. Relying solely on detection, moderation, or transparency is insufficient. Fact-checking and debunking, while essential, are insufficient tools to counteract the phenomenon. To protect democratic discourse, society must combine technological understanding with critical literacy, legal safeguards, and institutional resilience. By doing so, we can navigate the tension between innovation and stability, ensuring that synthetic media becomes a tool for creativity and pluralism rather than a persistent threat.

Chapter 1, "Unmasking the Illusion: The Tech Behind Deepfakes," introduces the technical foundations of synthetic media. It traces the development of artificial intelligence from early rule-based systems to generative models such as Generative Adversarial Networks (GANs) and diffusion models, explaining how advances in computational power and accessibility have transformed deepfake creation from a specialist activity into a widespread practice.

Chapter 2, "The Spread of Deepfakes in Digital Networks," examines how synthetic media circulate across social platforms. It explores the mechanisms of algorithmic amplification, virality, and emotional engagement, showing how networked structures reward deceptive or sensational content. The chapter also presents early detection approaches developed in the SOLARIS project, combining statistical modelling with sentiment analysis to identify emerging disinformation patterns.

Chapter 3, "Semiotics of Synthetic Media," approaches deepfakes as cultural texts rather than purely technical artifacts. It employs semiotic and Actor-Network Theory frameworks to analyse how meaning is produced through human and technological interactions. Drawing on examples such as the "Pope Balenciaga" image and the digital recreation of Dalida, the chapter illustrates how synthetic visuals challenge the viewer's assumptions about authenticity and representation.

Chapter 4, "The Psychology of Deception: Why We Believe Deepfakes," explores the cognitive and emotional processes that shape belief in synthetic content. It explains why individuals often fail to detect deepfakes and how factors such as ideology, prior knowledge, and media literacy influence vulnerability. The chapter introduces the Perceived Deepfake Trustworthiness Questionnaire (PDTQ) as a framework for understanding how presentation quality and content plausibility jointly affect belief and behavioural intention.

Chapter 5, "Democracy Distorted: Deepfakes as Political Weapons," examines how synthetic media has become a tool for political manipulation, transforming deepfakes from isolated forgeries into structural threats to democratic integrity. The chapter analyses how generative AI technologies enable the rapid and inexpensive creation of deceptive political content, eroding the evidentiary foundations on which public trust and accountability depend. Through case studies from Europe and the United States, it explores how deepfakes intensify epistemic and political harms, targeting vulnerable groups and undermining electoral processes. Drawing on political theory, the authors argue that deepfakes expose deeper weaknesses in the contemporary information ecosystem and call for institutional, educational, and policy measures capable of restoring transparency and civic trust.

Chapter 6, "Synthetic Media for Social Good: Unlocking Positive Potential," shifts the focus from risk to opportunity. It explores how deepfakes can be used ethically to support education, cultural preservation, and civic participation, aligning with the broader vision of "AI for Social Good." It also presents interpretive taxonomies developed through participatory research to evaluate how audiences perceive and assess positive synthetic content.

Chapter 7, "Governing Deepfakes: Legal Initiatives and Regulatory Gaps," provides a detailed examination of the current European legal landscape. It evaluates how the GDPR, the Digital Services Act, and the AI Act interact in addressing synthetic media and identifies ongoing gaps, such as the limited protection against immaterial harms and the persistent lag between technological and legislative developments.

Chapter 8, "Regulatory Innovations and Policy Options for Synthetic Media and Digital Democracy," concludes the book by outlining strategies for democratic resilience. It proposes unified personality rights, harmonized regulatory frameworks, and investments in media and AI literacy as essential to preserving trust in digital communication. The chapter emphasizes the need to move beyond reactive content moderation toward proactive policies that support pluralism, critical engagement, and ethical innovation.

The Conclusions bring together the insights developed throughout the book and outline future directions for research, policy, and ethical practice in addressing the evolving challenges of synthetic media.

End notes

All authors equally helped conceptualize, write, and edit the introduction.

UNMASKING THE ILLUSION: THE TECH BEHIND DEEPFAKES

MICHELE BRIENZA,[1] DOMENICO DANIELE BLOISI,[2] DANIELE NARDI[1]

[1] Sapienza University of Rome, Department of Computer, Control and Management Engineering "A. Ruberti", Rome, Italy
brienza@diag.uniroma1.it, nardi@diag.uniroma1.it
[2] International University of Rome - UNINT, Department of International Humanities and Social Sciences, Rome, Italy
domenico.bloisi@unint.eu

The chapter provides an overview of the technology behind deepfakes, describing what a deepfake is and how it is created. The chapter is structured around three sections: (i) theoretical foundations of artificial intelligence, machine learning, and deep learning, (ii) generative models and synthetic data, and (iii) the synthetic media toolkit. Firstly, it describes AI evolution, starting from the early stages leading up to the latest models that can generate data. The latest models are then described, highlighting their capabilities and explaining how these models open a wide range of opportunities, as well as the concerns regarding the generation of highly realistic data that can deceive users, as is the case with deepfakes. Finally, knowledge of how the machines learn from the data helps in using these tools. A clear understanding of the process behind the technology leads to unmasking the illusion and understanding how the technology works, enabling informed use.

DOI
https://doi.org/
10.18690/um.feri.2.2026.1

ISBN
978-961-299-109-8

Keywords:
Generative AI,
GANs,
deepfakes,
machine learning,
AI history

University of Maribor Press

1 Theoretical Foundations: Artificial Intelligence, Machine Learning, Deep Learning

A "deepfake" is a digital content (i.e., an image, a video, or an audio) modified or generated by using artificial intelligence (AI) tools. The term combines two concepts: "deep" refers to deep learning, a branch of AI, while "fake" indicates that the content has been manipulated or altered in some way. So, to understand deepfakes, it is important to first examine the theoretical foundations of AI.

The objective of AI is to create machines capable of performing tasks that typically require human intelligence. The journey begins in the 1950s, when pioneers such as Alan Turing posed the fundamental question of machine intelligence: "Can machines think?" (Epstein et al. 2009). This chapter provides a brief technical overview of how the technology born in 1950 has evolved to today's capability of creating highly realistic synthetic data content.

The path from Turing's question to today's deepfakes not only regards technological advancement, but it is also a fundamental transition in how machines process and generate information. In scientific evolution over the last decades, we have seen a progression from rule-based systems to machine learning and, finally, to deep neural networks. This evolution has led nowadays to models that can generate realistic human faces, synthesize speech in any voice, and produce entirely fictional yet photorealistic scenarios.

The formal birth of the term "artificial intelligence" took place in 1956, when, during a conference, John McCarthy, one of the pioneers of AI, coined the term (McCarthy et al. 2006). However, the conceptual foundations were laid earlier by Alan Turing, whose 1950 paper "Computing Machinery and Intelligence" introduced the famous Turing Test as a benchmark for machine intelligence, posing the question, "Can machines think?"

Developing a machine capable of "thinking" is a challenge that extends beyond mere technical complexity. It requires sophisticated models, advanced computational architectures and, crucially, a profound sense of responsibility. The aim is to create systems that can produce high-quality content while adhering to fundamental ethical principles; a balance that is becoming increasingly important as these technologies become more powerful and widespread.

To understand how machines learn to replicate human intelligence, Tom Mitchell provides a foundational definition of the learning process: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997).

Early AI development focused on symbolic AI or knowledge-based systems. These systems operated on the principle that intelligence could be replicated through explicit rules and logical reasoning. Engineers would interview domain experts, codify their knowledge into if-then rules, and create systems that could make decisions within structured domains. Due to their rule-based architecture, small changes in input could lead to system failures; to operate in a human-like fashion, a machine requires learning from experience or adapting to new situations not explicitly programmed. Consider image recognition: the exclusive use of standard geometric features such as size, colour, and pose is not sufficient for this task, which requires a more robust approach. Machine learning has emerged as a solution to these limitations, representing a significant advance in the field of artificial intelligence. Instead of programming explicit rules, ML enables systems to learn patterns from data. The fundamental idea is that human behaviours can be learned through data without being specified by a set of rules.

Machine learning comprises three main paradigms:

- **Supervised Learning**: a machine learning paradigm that involves training algorithms on labelled datasets, where each training example consists of an input paired with its corresponding correct output (label). The system learns to map inputs to desired outputs by analysing these input-output pairs during the training phase. Through this process, the algorithm identifies patterns and relationships within the data that enable it to make accurate predictions. Classification tasks (determining whether an email is spam) and regression problems (predicting house prices) are examples of supervised learning. The mathematical foundation rests on finding functions that minimize prediction errors across training data while generalizing well to unseen examples.

- **Unsupervised Learning**: a machine learning paradigm that addresses the challenge of discovering hidden patterns, structures, and relationships within data without the guidance of explicit labels or target outputs. Unlike supervised learning, these algorithms must identify meaningful patterns just using the input data itself, making this approach particularly valuable for exploratory data analysis and knowledge discovery. These algorithms might cluster customers into market segments, reduce data dimensionality for visualization, or discover anomalies in network traffic. Principal Component Analysis (PCA) and k-means clustering represent classical unsupervised techniques that remain widely used today.
- **Reinforcement Learning**: a machine learning paradigm that draws inspiration from behavioural psychology, where agents learn through trial and error in interactive environments. The agent receives rewards or penalties for actions, gradually learning optimal strategies to reach a specified goal. This approach has produced remarkable successes in game-playing AI, from chess programs to AlphaGo's historic victory over human champions.

The early 2000s marked a pivotal transformation in artificial intelligence, driven by exponential growth in computational resources and unprecedented access to large datasets. This technological convergence enabled the emergence of Deep Learning methodologies that fundamentally changed how machines process information. Complex neural architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) demonstrated remarkable capabilities in pattern recognition, language understanding, and cross-modal translation tasks that had previously been challenging. These networks are made of layers that learn patterns from input data to predict outputs.

A salient moment arrived in 2014, with Ian Goodfellow's introduction of Generative Adversarial Networks (GANs), establishing the foundation for modern generative artificial intelligence (Goodfellow et al. 2014). This breakthrough represented more than incremental progress; it introduced an entirely new paradigm for synthetic data creation.

The GAN framework operates through an adversarial training process involving two competing neural networks. The generator network learns to transform random noise into increasingly synthetic content, whether images, audio, or text. Meanwhile, the discriminator network distinguishes whether the content in the input is generated or not, acting as a digital detective. This competitive process creates a feedback loop where both networks continuously improve: the generator becomes more skilled at creating convincing fakes, while the discriminator becomes better at detecting them. Eventually, this adversarial process reaches an equilibrium where generated content achieves a high level of realism.

However, the most revolutionary advancement in generative AI came with the 2017 introduction of the Transformer architecture (Vaswani et al. 2017), which transformed both natural language processing and artificial intelligence. Transformers introduced the attention mechanism, enabling models to dynamically focus on relevant input segments during processing. This architecture consists of an encoder that builds rich contextual representations through self-attention, and a decoder that generates outputs autoregressively by considering both encoded input and previously generated tokens. This innovation directly enabled the development of GPT models by OpenAI, which demonstrated unprecedented natural language capabilities through large-scale pre-training on diverse text corpora, fundamentally changing how machines understand and generate human language

The impact of GANs has led to advantages across numerous applications. Beyond generating photorealistic synthetic faces of non-existent individuals, these models have revolutionized creative industries, enhanced image super-resolution techniques, and provided synthetic training data solutions when real datasets are limited or prohibitively expensive. GANs essentially established the conceptual groundwork for the generative AI revolution, inspiring subsequent innovations including diffusion models and transformer-based architectures.

While GANs dominated early generative AI development, diffusion models (Ho et al. 2020). emerged as an alternative, particularly for image synthesis tasks. These models employ a fundamentally different approach: rather than direct generation, they learn to progressively remove noise from random input through iterative refinement steps. This denoising process offers greater training stability and finer control over the generation process. Contemporary models like DALL-E (Ramesh et al. 2021) and Stable Diffusion (Rombach et al. 2022) exemplify how diffusion approaches can produce both artistic and photorealistic imagery with unprecedented precision and creative flexibility.

2 Generative Models and Synthetic Data

The original GAN, called Vanilla GAN (Goodfellow et al. 2014), marked the beginning of adversarial networks. Training GANs is challenging due to technical issues such as mode collapse and unstable gradients that cause not high-quality output generation, but this model laid the groundwork for the future development of GAN architectures. Since the introduction of Vanilla GANs, numerous improvements and variations have been proposed to address limitations and expand the applicability of GANs. Some of the major advancements include:

- **Conditional GANs (cGANs)** (Mirza and Osindero 2014) introduced the ability to condition the generation process on additional information, such as class labels. By conditioning both the generator and discriminator on a desired output class, cGANs allowed for greater control over the generated outputs. This was a critical advancement in applications like image-to-image translation and text-to-image generation.
- **Deep Convolutional GAN (DCGAN)** (Radford et al. 2016) leveraged convolutional neural networks (CNNs) to improve the stability and quality of generated images. DCGANs enabled the generation of more detailed and higher-resolution images by using convolutional layers in both the generator and discriminator. This model became a benchmark for image synthesis tasks and paved the way for many subsequent GAN models.
- **Wasserstein GAN (WGAN)** (Arjovsky et al. 2017) addressed the training instability issue by using the Wasserstein distance to measure the difference between real and generated data distributions. This led to more stable training and better convergence.
- **Progressive GAN (PGAN)** (Karras et al. 2018) improved the generation of high-resolution images by starting with a low-resolution image and progressively increasing the resolution during training. This method helped generate more stable and realistic images.
- **CycleGAN** (Zhu et al. 2020) introduced the concept of cycle consistency, enabling unpaired image-to-image translation. This meant that CycleGANs could learn to transform images from one domain to another without needing paired training data, making it highly versatile for applications like photo enhancement and style transfer.

- **StyleGAN** (Karras et al. 2019) introduced a new way to control the generation process by manipulating latent space features to produce human faces. StyleGAN's ability to disentangle features like pose, facial expression, and hairstyle in the generation process resulted in highly realistic images. StyleGAN2 (Karras et al. 2020) and StyleGAN3 (Karras et al. 2021) were followed by improvements, including better handling of textures and the ability to create more coherent images across different resolutions.
- **BigGAN** (Brock et al. 2019) enhanced the performance of GANs by training on large-scale datasets with higher computational power. BigGANs demonstrated the capability of GANs to generate incredibly detailed and high-quality images, pushing the boundaries of what GANs could achieve.
- **Self-Attention GAN (SAGAN)** (Zhang et al. 2018) introduced self-attention mechanisms that allowed the network to focus on relevant parts of an image while generating it. This enabled the generation of images with greater structural and spatial consistency.
- **DragGAN** (Pan et al. 2023) represents a novel approach to interactive image manipulation. It enables users to control specific points in an image and drag them to target positions, providing precise control over shape, pose, and expression. This interactive manipulation method enhances user control in generating and editing images.

The evolution of GANs from simple image generation to sophisticated manipulation tools marks a critical turning point in synthetic media creation. As these models became more powerful and accessible, they enabled a new phenomenon that would capture attention: deepfakes. Thanks to their ability to generate and modify existing multimedia content with increasingly realistic results, the phenomenon of deepfakes has spread widely. Deepfakes are synthetic content created through sophisticated artificial intelligence models, particularly GANs and diffusion models, which allow for convincing manipulation of videos, images, and audio recordings.

This technological capability, while impressive from an engineering perspective, has introduced unprecedented challenges. The same algorithms that can enhance medical imaging or create innovative art can also fabricate convincing videos of public figures saying things they never said, or place individuals in scenarios they never experienced. The democratization of these tools has made synthetic media creation accessible beyond academic laboratories, raising fundamental questions about truth, authenticity, and the nature of evidence in our digital age.

3 Beyond GANs: The New Wave of Generative Models

While GANs have dominated the field of image synthesis and related tasks for almost a decade, other approaches have shown promising results in producing highly detailed and controllable outputs. These alternative models have provided new perspectives on how generative AI can be approached.

Variational Autoencoders, introduced by Kingma and Welling in 2013 (Kingma and Welling 2013), represent one of the earliest and most influential alternatives to GANs in the generative modelling landscape. VAEs combine the concepts of autoencoders with variational inference, creating a probabilistic framework for learning latent representations of data. The architecture consists of two main components: an encoder network that maps input data to a probabilistic latent space, and a decoder network that reconstructs the original data from latent representations. Unlike traditional autoencoders that learn deterministic mappings, VAEs learn to encode data into probability distributions in the latent space, typically Gaussian distributions characterized by mean and variance parameters.

The key advantage of VAEs lies in their stable training process, which leads to more stable and predictable training compared to the adversarial training of GANs. The probabilistic nature of VAE latent spaces enables smooth interpolations between different data points, making them particularly useful for tasks requiring controlled generation and data exploration. However, VAEs also have notable limitations, particularly in generating sharp, high-resolution images, where they tend to produce somewhat blurry outputs compared to GANs.

The introduction of the Transformer architecture (Vaswani et al. 2017) revolutionized natural language processing and opened new possibilities for generative modelling across multiple modalities. Originally designed for sequence-to-sequence tasks, Transformers have proven to be remarkably versatile and powerful for generative applications. The self-attention mechanism allows Transformers to model long-range dependencies effectively, making them particularly suitable for sequential data generation. Unlike classic methods such as RNNs or CNNs, transformers can process entire sequences in parallel and capture complex relationships between distant elements.

Most transformer-based generative models work autoregressively, predicting the next token in a sequence given all previous tokens. This approach has proven highly effective for text, code, and even image generation when images are treated as sequences of tokens. The Generative Pre-Trained Transformer (GPT models) are language models. These models showed that scaling up transformers with massive datasets could lead to emergent capabilities in text generation, reasoning, and even multimodal understanding.

Building upon the foundations laid by VAEs and the architectural innovations of Transformers, Diffusion Models have emerged as perhaps the most significant advancement in generative AI in recent years. These models have achieved unprecedented quality in image generation and are rapidly expanding to other modalities. Diffusion models work by modelling the process of data generation as the reverse of a diffusion process, starting with random noise and gradually denoising it through a series of iterative steps to generate data that resembles the original distribution.

The generation quality of diffusion models has been demonstrated through their exceptional capability in generating highly detailed, realistic images that often surpass the quality of both GAN and VAE-generated content. Unlike GANs, diffusion models do not suffer from adversarial training instabilities, and unlike early transformer approaches, they do not require massive computational resources for basic functionality. These models can cover the data distribution more accurately than GANs, allowing them to generate a wider variety of high-quality content.

4 The Deepfake Pipeline: Tools and Techniques for Synthetic Content Creation

Deepfakes, a specific application of GANs, have become a key technology for generating hyper-realistic fake videos and audio. Deepfakes allow for the alteration of visual and auditory content in a manner that is nearly indistinguishable from real media. While the foundational models and applications mentioned here represent the pioneering technologies that first brought deepfakes to mainstream attention, the field is currently experiencing unprecedented rapid evolution. Early tools like **FakeApp**, **FaceSwap**, and **Face2Face** established the fundamental principles, but today's landscape features increasingly sophisticated architectures, real-time processing capabilities, and improved accessibility. The focus has shifted from experimental proof-of-concepts to robust, production-ready toolkits that can deliver professional-quality results with minimal technical expertise required from users. Below are the major techniques used in deepfake generation and their current implementation frameworks:

- **Face-swap technology** is the most recognized form of deepfake. It involves replacing a person's face in a target video with the face of another individual by training an AI model, one of the ones seen in the previous section, on two sets of facial images: the source face and the target face. The model learns to encode the distinctive features of the source person's face into a latent representation, then reprojects these encoded features onto the target person's facial structure. The face swap process allows for automatic swapping of facial features while maintaining the context and integrity of the original video's environment, adjusting for different face shapes, angles, lighting conditions, and camera perspectives.
- **Lip-syncing deepfakes** manipulate the movement of the lips in a video to match a specific audio input. This technique takes the target video frames and the desired audio as inputs, analysing the phonetic structure and temporal patterns of the speech to generate corresponding visual mouth shapes and facial muscle movements. The model encodes the audio features and learns the correlation between speech sounds and their visual representations, generating a video where the target person's mouth movements are synchronized with arbitrary speech audio. Advanced models employ GenAI to generate real-time lip movements, eye-blinking, and facial expressions that naturally accompany speech, resulting in highly realistic video content.
- **Face reenactment** is a deepfake technique where a source actor's facial expressions, gestures, and head movements are transferred to a target video. The system takes video input from both a source performer and a target person, encoding the source's facial dynamics, including expression parameters, head pose, and micro-movements, into a control representation. This encoded motion

data is then reprojected onto the target's facial structure and appearance, enabling the modification of a person's facial expressions in real-time or recorded videos. This makes it appear as if they are displaying emotions or gestures, they never performed, while preserving the target person's identity and the original video context.

- **Voice cloning** generates synthetic speech that mimics a target person's voice characteristics and speaking patterns. This technique utilizes pre-trained models that can clone voices with minimal input requirements, requiring only a few seconds of audio samples from the target speaker. The system takes the reference audio sample and the desired text as inputs, encodes the vocal characteristics from the audio sample, and generates new speech content that maintains the original speaker's voice while saying the provided text. These models can reproduce acoustic properties, speaking mannerisms, and natural speech variations, enabling the generation of convincing audio content where the target appears to be saying words or phrases they never actually spoke.

The evolution of deepfake technology from academic research to practical applications has followed a predictable pattern of democratization. Initially dominated by complex open-source frameworks requiring substantial technical expertise, the field has progressively become more accessible through lighter models and, most recently, commercial platforms that abstract away all technical complexity.

The early deepfake ecosystem was built around open-source projects. DeepFaceLab emerged as the dominant force, responsible for creating the majority of professional deepfake content. This comprehensive framework provided end-to-end functionality for face extraction, training, and merging, but demanded significant technical knowledge and powerful hardware. Users needed to understand neural network architectures, manage training parameters, and navigate complex file structures. FaceSwap offered a similarly powerful alternative with better cross-platform support and multi-GPU capabilities, built on TensorFlow and Python.

This technical and financial barrier initially limited the creation of deepfakes mainly to academic researchers and industry professionals who used the technology for academic studies or private work. However, demonstrations of the technology's capabilities quickly attracted the interest of a wider community, including people with malicious intentions. Online forums and GitHub communities began sharing code, tutorials, and pre-trained models. The availability of detailed guides and community support made the technology increasingly accessible, although it still required considerable technical expertise and a significant hardware investment. Cloud platforms played a crucial role in this democratization. Google Colab, launched in 2017 as a tool for research and training in the field of machine learning, offered free access to powerful GPUs. This allowed anyone to experiment with deep learning models without having to purchase expensive hardware by following online guides. Although many have used these resources for legitimate projects, unfortunately, some uses have involved misuse for the creation of non-consensual deepfakes. To circumvent the problem in 2022, Google implemented specific restrictions in its terms of service, explicitly prohibiting the use of Colab for the generation of deepfakes.

The landscape began shifting with the development of lighter, more efficient models that could run on consumer hardware. These optimized architectures reduced training times and memory requirements, making deepfake creation feasible on standard gaming PCs. Real-time applications like DeepFaceLive demonstrated that face swapping could be performed live during video calls or streaming, achieving real-time performances by using a common laptop.

Today's synthetic media landscape has been transformed by commercial platforms that have eliminated virtually all technical barriers. It is possible to generate multimedia content instantly through services offered to users on websites. Services such as FakeYou and DeepSwap offer professional-quality results with a few simple clicks. These platforms handle all the computational complexity in the cloud, allowing users to create convincing multimedia content simply by uploading images and videos. Subscription models typically start at tens of dollars per month, making the technology accessible to anyone willing to pay the cost of the service to generate content that can be used for entertainment purposes.

Even large companies have released their own models for using video generation models for benevolent purposes. HeyGen, Veo3, and Runway are some of the models used to generate videos useful for advertising campaigns or simply for fun.

The above-mentioned models were then developed by design to be used ethically to avoid the generation of malicious content. The landscape of video generation has been transformed by models such as Runway and Google's Veo 3, which represent significant advances in both quality and control. Runway Gen-4 introduces consistent character generation across scenes, allowing filmmakers to maintain visual continuity while generating content from multiple perspectives. The model excels at realistic physical simulation and can use visual references combined with textual instructions without requiring fine-tuning. Google's Veo 3 goes a step further by generating videos with native audio integration, including synchronized dialogue, ambient sounds, and music, creating 8-second clips that achieve cinematic-level realism while maintaining built-in security protocols.

The development of these platforms has opened up a world of opportunities for advertising campaigns, corporate training, multilingual customer support, and educational content thanks to software such as HeyGen, which focuses on creating AI avatars for business applications such as The platform can create highly realistic digital twins from a single photo, with advanced understanding of the script that regulates facial expressions, body language, and voice inflections to match the meaning of the content, while maintaining built-in protections and human moderation to prevent misuse.

The synthetic media toolkit has been further expanded by advanced image manipulation capabilities. Google's Nano Banana, integrated into Gemini, represents a breakthrough in natural language-based image editing that transforms how synthetic content is created. Unlike traditional photo editing software requiring technical skills, Nano Banana allows users to modify images through simple conversational prompts. Users can seamlessly blend multiple photos, change backgrounds, alter clothing and appearance, or place subjects in entirely new environments while maintaining photorealistic consistency. The model excels at character preservation, ensuring that people and animals retain their distinctive features across edits, making it particularly powerful for creating convincing synthetic scenarios.

This capability transforms content creation workflows by eliminating the traditional barrier between imagination and execution. Content creators can now generate complex composite images by describing desired changes rather than mastering complex editing techniques. The model's ability to maintain spatial coherence and lighting consistency across edits makes it possible to create highly believable synthetic content that would previously require professional photography and post-production skills.

These advanced platforms incorporate sophisticated content moderation systems designed to prevent harmful applications. The model includes embedded content filters that restrict inappropriate content generation, visual watermarking with SynthID for authenticity verification, and metadata identification to maintain provenance tracking. This "safe-by-design" philosophy implements restrictions during the development process rather than relying solely on post-generation filtering.

5 Concluding Remarks

This chapter examined the technological foundations of artificial intelligence that have taken this discipline from rule-based systems designed to mimic human intelligence to modern generative models capable of generating synthetic media, representing a fundamental change in the way multimedia content is generated.

The generative artificial intelligence ecosystem has spread very quickly, starting with GANs, which, with their variants, have led to the development of models such as StyleGAN, capable of generating non-existent human faces with a very high level of quality, and transformers for textual and multimodal applications. An example of the use of these models is GPT and all the large language models currently in use, which allow the generation of very high-quality text and are capable of generating text documents with human-like "reasoning" capabilities. Then there are diffusion models, which achieve image generation and editing with unprecedented image quality through iterative denoising.

It is important to note that the development of these models requires huge data sets, high-performance GPU clusters, and weeks of training. The technology has evolved from resource-intensive research projects to cloud-based services that hide all the technical complexity, transforming the creation of synthetic content from an exclusive research domain to a tool. The technology has evolved from resource-intensive research projects to cloud-based services that hide all the technical complexity, transforming the creation of synthetic content from an exclusive research domain to a tool accessible to everyone. This is a positive aspect in terms of the democratization of technology, but it raises ethical risks to consider regarding the spread of inauthentic content that can mislead users or defame individuals.

For this reason, the models developed by companies have built-in security measures that include content filters that prevent the generation of harmful content, training data curation, watermarking technologies such as SynthID, and API-level restrictions. Although imperfect, these represent an improvement over the early unrestricted models, which, in the early stages of generative artificial intelligence, allowed non-consensual or defamatory content to be spread online.

The fundamental challenge is that algorithms that enable legitimate applications in entertainment and education can equally serve harmful purposes, such as disinformation and fraud.

End notes

Michele Brienza is the main author of this chapter. He wrote all the sections that briefly describe the history of AI, from its birth to new generative AI models, and the tools and processes that enable the creation of synthetic media. Domenico Daniele Bloisi and Daniele Nardi collaborated in the organization of the content and final review of the chapter.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. *Proceedings of the 34th International Conference on Machine Learning*,214–223. https://proceedings.mlr.press/v70/arjovsky17a.html

Brock, A., Donahue, J., & Simonyan, K. (2018, September 27). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *International Conference on Learning Representations.* https://openreview.net/forum?id=B1xsqj09Fm

Epstein, R., Roberts, G., & Beber, G. (Eds.). (2009). Parsing the Turing Test: Philosophical and Methodological Issues, *Quest for the Thinking Computer*. Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances, *Neural Information Processing Systems*, 27. https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf

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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation, *arXiv*. https://doi.org/10.48550/arXiv.1710.10196

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-Free Generative Adversarial Networks. Advances, *Neural Information Processing Systems*, 34, 852–863. https://proceedings.neurips.cc/paper/2021/hash/076ccd93ad68be51f23707988e934906-Abstract.html

Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 4396–4405. https://doi.org/10.1109/CVPR.2019.00453

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 8110–8119. https://openaccess.thecvf.com/content_CVPR_2020/papers/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.pdf

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

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. https://doi.org/10.1609/aimag.v27i4.1904

Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets, *arXiv*. https://doi.org/10.48550/arXiv.1411.1784

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., & Theobalt, C. (2023). Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. A*CM SIGGRAPH 2023 Conference Proceedings*, 1–11.https://doi.org/10.1145/3588432.3591500

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,arXiv. https://doi.org/10.48550/arXiv.1511.06434

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. https://doi.org/10.1109/CVPR52688.2022.01042

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

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-Attention Generative Adversarial Networks. *Proceedings of the 36th International Conference on Machine Learning*, 7354–7363. https://proceedings.mlr.press/v97/zhang19d.html

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244

THE SPREAD OF DEEPFAKES IN DIGITAL NETWORKS

TOMMASO TONELLO,[1] ASENIYA DIMITROVA,[2]
LIVIO FENGA,[3] LUCA BIAZZO,[4] ALESSIO JACONA[5]

[1] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
t.tonello@uu.nl
[2] Brand Media Bulgaria, Sofia, Bulgaria
a.dimitrova@economic.bg
[3] University of Exeter, Department of Management, Exeter, United Kingdom of Great Britain and Northern Ireland
l.fenga@exeter.ac.uk
[4] University of Brescia, Department of Economics and Management, Brescia, Italy
luca.biazzo@unibs.it
[5] Agenzia Nazionale Stampa Associata, Rome, Italy
alessio.jacona@gmail.com

DOI
https://doi.org/10.18690/um.feri.2.2026.2

ISBN
978-961-299-109-8

This chapter explores the spread of deepfakes through social media platforms (e.g. X, Facebook and TikTok). By studying real-world case studies, such as political deepfakes or celebrity impersonations, the chapter illustrates how synthetic media exploit online engagement dynamics to reach massive audiences quickly. It then reviews current methods used to detect and track deepfakes, especially early-warning systems monitoring content spread patterns to flag potential deepfakes in real time, as well as novel research instruments developed as part of the SOLARIS project. The chapter then presents the role of traditional media in debunking and contextualising deepfakes, reflecting upon the challenges that AI-generated disinformation poses to journalists and media professionals. In this context, insights from SOLARIS' Use Case 2 are used to show how targeted interventions can slow the spread of harmful synthetic media. Finally, the chapter advocates for bottom-up AI education to frame digital citizens' needs and to foster their ability to engage with online synthetic content.

Keywords:
deepfakes,
social media,
disinformation,
statistical analysis,
journalism,
education

University of Maribor Press

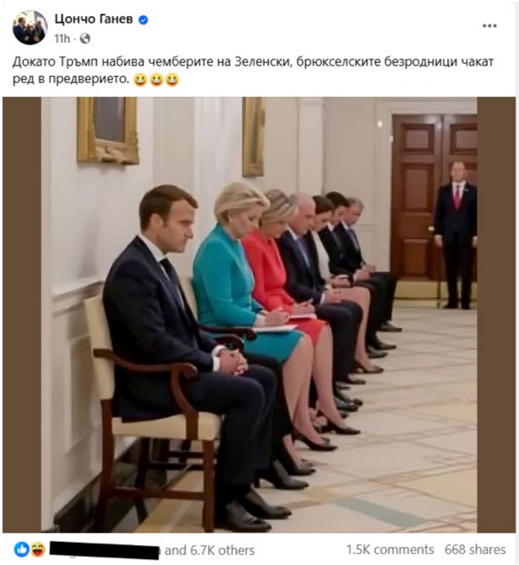
1 The Problem of Deepfakes and Social Media: How Deepfakes Go Viral

In an age where digital content moves at unprecedented speed, deepfakes have emerged as one of the most disruptive forms of synthetic media. Their increasing realism and accessibility raise pressing concerns about the manipulation of public opinion and democratic engagement, especially in politically sensitive contexts. This chapter investigates how deepfakes propagate across digital networks, with a particular focus on the architecture of platforms such as X (formerly Twitter) and Facebook. These environments, governed by engagement-driven algorithms and virality incentives, are especially susceptible to the rapid diffusion of deceptive content. Understanding these dynamics is essential to anticipating, detecting, and ultimately mitigating the societal risks posed by deepfakes. The analysed cases offer insights into how disinformation is packaged for viral spread. Ultimately, we point to the need for cross-disciplinary approaches, combining technical detection, network modelling, social media analysis, and media experts' insights, to map and counteract the spread of deepfakes and to disseminate relevant AI knowledge at the societal level.

This section details how deepfakes go viral on social media, drawing on examples from the U.S. and European political landscapes. The examples were picked based on their prominence and recency. It is beyond the scope of this chapter to analyse all cases and uses of disinformation using AI-generated content. Instead, we picked five examples that left a mark by reaching large audiences. Each of them is illustrative of the use of different social media channels based on the goals of misleading posts created or shared by online users. We then draw some conclusions on the mechanisms by which online network algorithms can enhance deepfake distribution.

1.1 The Case for the Obedient European Leaders

A most recent example comes from the context of the Russia-Ukraine war. It follows a meeting between European leaders at the White House on 18 August 2025, which took place as part of the peace-building efforts by the Trump administration. During the event, President Trump had lengthy discussions with Western leaders, including French President Emmanuel Macron and the European Commission President Ursula Von der Leyen, to agree on a common negotiating position.



Цончо Ганев
11h ·

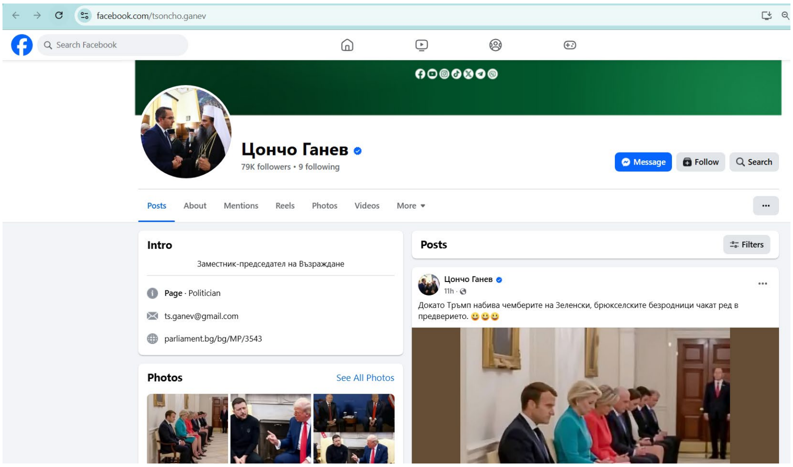
Докато Тръмп набива чемберите на Зеленски, брюкселските безродници чакат ред в предверието. 😃😃😃

👍😆 [redacted] and 6.7K others 1.5K comments 668 shares

Figure 2.1: Detail of European leaders queuing to meet President Trump (deepfake)
Source: https://www.facebook.com/tsoncho.ganev.

However, a widely circulating deepfake image claimed to be taken on the day of the event portrays European leaders sitting obediently, heads down, waiting for the American President to return with instructions. The post claims to show the leaders waiting for President Trump to finish schooling President Zelensky. Whatever the interpretation, the message is clear: European politicians are portrayed as showing weakness, being sidelined by the great leaders of Russia and the USA, and they are only observers of important events in international politics. Only, this never happened, and there are clear signs that the image is fake.

The image has been circulating widely on Facebook and X, but the screenshot showcased above is taken from the Facebook page of a high-ranking pro-Russian politician from Bulgaria, Tsoncho Ganev, member of Parliament and vice-president of the pro-Russian Vazrazhdane (Revival) party, which maintains close ties to Putin's United Russia party, the two having recently signed a collaboration agreement. The post caption reads in slang: While Trump is schooling Zelensky, the barren Brusselers are waiting their turn in the lobby. Ganev is probably not the real author of the image, because the quality is low (suggesting he probably saw it somewhere and took a screenshot). Nevertheless, he started an information thread which became widespread in Bulgaria.



Цончо Ганев

79K followers • 9 following

Intro

Заместник-председател на Възраждане

Page · Politician

ts.ganev@gmail.com

parliament.bg/bg/MP/3543

Photos

Posts

Цончо Ганев

11h ·

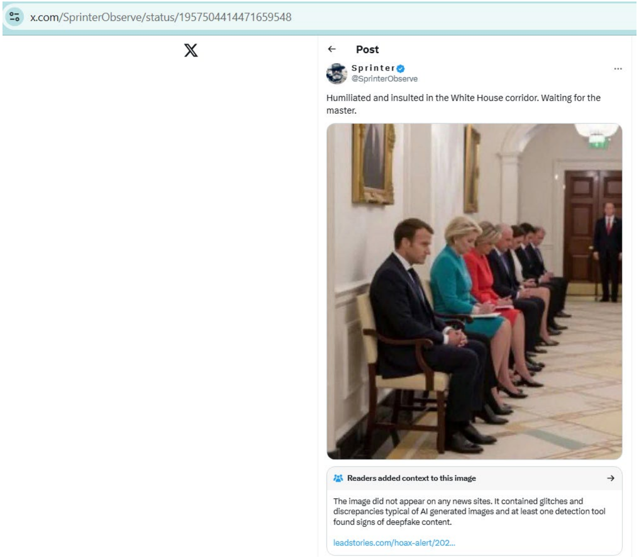
Докато Тръмп набива чемберите на Зеленски, брюкселските безродници чакат ред в предверието. 😀😀😀

Figure 2.2: European leaders queuing to meet President Trump (whole post on Facebook)
Source: https://www.facebook.com/tsoncho.ganev

Within hours, the image had amassed hundreds of shares and thousands of comments, mostly supportive, although it clearly is a deepfake – this can be seen by several inconsistencies, including a pair of legs with no body between the French and the EU Commission presidents, a mismatch between the outfits they actually wore that day and those shown in the image, a difference between President Macron's shoes, etc. At the time of writing, and despite several reports to Facebook that the image is false, it has not been taken down, nor has any context been added by the network to label it as a deepfake. The same politician has also shared the content on their X page, but since this network is not highly popular in Bulgaria, the effect it produced there was of a different magnitude.

A basic manual review of shares shows that among the profiles that have shared the image on Facebook, there are genuine profiles, largely pro-Russian supporters, official pages of political party structures, and many fake profiles (with fake images, low numbers of friends, mostly propaganda-style content). The post has also been shared in several Facebook groups publishing, among other things, anti-Western integration (for example, one that opposes Bulgaria's integration into the Eurozone), anti-establishment, and anti-George Soros content. This testifies to the importance of information bubbles on social media, safe spaces where we encounter mostly information that fits into our own worldviews and comes from sources that we consider safe and credible.

Meanwhile, on X, the same image shared by a verified user (called Sprinter Observe), with 770 k+ followers, has accumulated 104.8K views within a few hours and a similar number of shares. However, we can already see readers having generated a contextual note, saying that this is a fake image and explaining why.



x.com/SprinterObserve/status/1957504414471659548

Sprinter
@SprinterObserve

Humiliated and insulted in the White House corridor. Waiting for the master.

Readers added context to this image

The image did not appear on any news sites. It contained glitches and discrepancies typical of AI generated images and at least one detection tool found signs of deepfake content.

leadstories.com/hoax-alert/202...

Figure 2.3: European leaders queuing to meet President Trump (X version of the post)
Source: https://www.facebook.com/tsoncho.ganev (Tsoncho Ganev on Facebook)

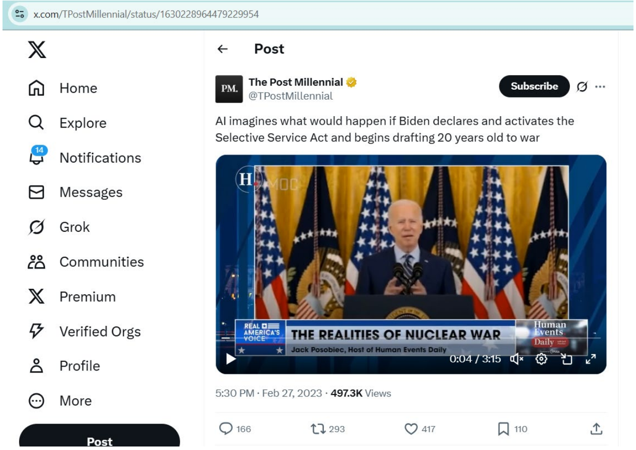
The post caption reads: "Humiliated and insulted in the White House corridor. Waiting for the master." This indicates that the author of the post intended to present it as true. While in the first example, the author of the post is clear, a known political figure aiming to strengthen their fan base and solidify support behind pro-Russian views in a critical time, in the X case, there is not much information about the author of the post. It claims to be an independent media reporter, but there is no additional public data to associate it with someone's identity. The only external link from the profile leads to a donation page. A reverse search of the profile image shows it is a portrait of Issam Zahreddine, one of the main commanders of Bashar al-Assad's army, killed in Syria in 2017, hence, not the real author of the post. This did not prevent the content from becoming viral, nor has it prompted the network to take down the profile or limit its exposure as being non-genuine.

This case might not be the most prominent example of social media use of deepfakes to harm, but it is pertinent and clearly shows the rapid spread of falsified content on Facebook, which can be re-shared with a lack of criticism and powered by influential figures from the political world and from the civic side itself. The fact that the posts have not been removed from their authors' profiles and no explanation for their authenticity has been given suggests that the intention has never been to inform, but rather to create a lasting impression. A large body of experimental literature shows that misinformation often continues to influence people even after it has been explicitly debunked - the so-called continued influence effect (Lewandowsky et al., 2012). Corrections reduce but frequently do not fully eliminate the influence of the false claim; in some circumstances, corrections can fail or even (rarely) backfire. Therefore, a falsified picture such as the example above would leave a lasting impression on the audience, and the longer it stays online, the stronger the impression. We simply cannot unsee a picture, even if we have later been made aware that it has been manipulated.

1.2 President Biden Calling for a National Draft to Defend Ukraine

Another interesting example, once again from the context of the war in Ukraine, comes in the form of a deepfake video circulated on X. It depicts then-President of the USA Joe Biden during a briefing calling for a national draft allowing for men and women from the States to be called to fight in Ukraine. One of the first appearances of this content occurs on X on 27 February 2023 by a news aggregator called The Post Millennial. The caption of the post clearly states that the video is AI-generated, only to depict a fictitious scenario. A commentator later in the video also confirms this is not a real event, but content that has been scripted and designed by the production. A more detailed check establishes that the new video was a doctored version of a video released by the White House on another occasion back in 2021.

The video is a relatively good deepfake, as it is somehow credible in the sense that it depicts something that many people feared might happen; the images and sound are also realistic, and only a deeper look into the gestures of Biden shows that something is wrong. The post has gathered a considerable number of views and shares, but it is nothing unusual, given that the page is a popular one with over 430k followers.



x.com/TPostMillennial/status/1630228964479229954

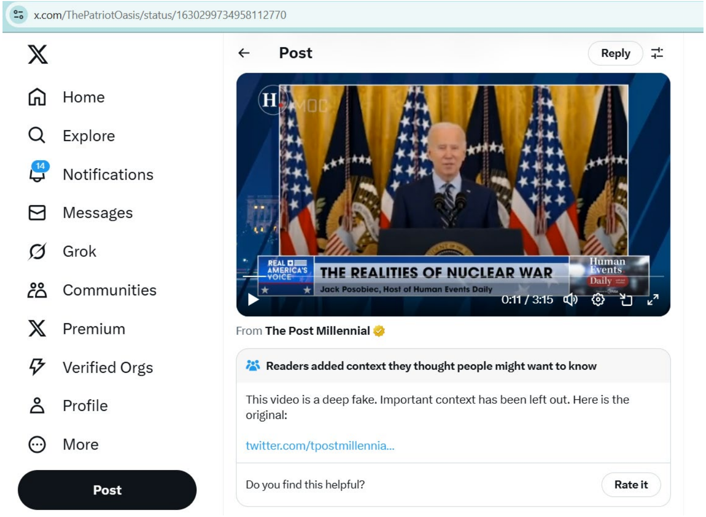
The Post Millennial @TPostMillennial

AI imagines what would happen if Biden declares and activates the Selective Service Act and begins drafting 20 years old to war

5:30 PM · Feb 27, 2023 · 497.3K Views

Figure 2.4: Video of President Biden announcing a national draft
Source: https://x.com/ThePatriotOasis/status/1630299734958112770.

The situation becomes much more interesting, as the video has been re-shared (although with a very different caption) by another page on X, the Patriot Oasis. While it has a smaller fan base than the original, the post has now accumulated over 8 million views. The difference: it presents the video as if it were genuine, using words such as "BREAKING" for a stronger emotional effect. The fact that it comes from a patriotic page might also have contributed to this.



x.com/ThePatriotOasis/status/1630299734958112770

From The Post Millennial

Readers added context they thought people might want to know

This video is a deep fake. Important context has been left out. Here is the original:

twitter.com/tpostmillennia...

Do you find this helpful?

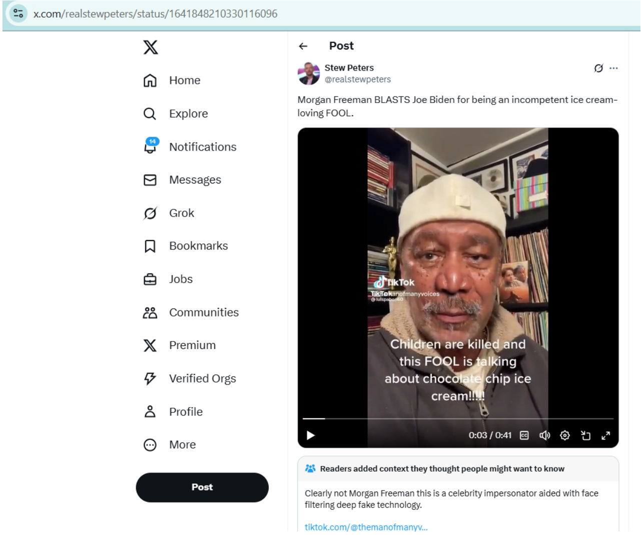
Figure 2.5: disclaimer flagging President Biden's national draft video as deepfake
Source: https://x.com/ThePatriotOasis/status/1630299734958112770.

This time, we can see that readers have added context explaining that the video is fake, but we do not know how many people noticed the warning and were influenced by it in the meantime. The different distributions of the same piece of content clearly show how disinformation spreads as fast as real news.

Growing scientific evidence shows that negative emotions, such as fear, anger, anxiety, and sadness, are systematically used on social media to amplify the spread of disinformation and, importantly, online engagement - to the benefit of social media platforms (Ali Adeeb & Mirhoseini, 2023).

1.3 Morgan Freeman calling President Biden a fool

Another example from the political sphere comes from the USA, but this time, there are two targets: the protagonist of the video, American actor Morgan Freeman, and Joe Biden, against whom the deepfake is addressed. The video depicts a poor-quality Freeman allegedly criticizing the President for being irrelevant in the situation of a mass shooting in the USA and calling for his removal from office. Originally, the video appeared on TikTok but was deleted: the post below comes from a repost of conservative radio host Stew Peters with the caption Morgan Freeman BLASTS Joe Biden for being an incompetent ice cream-loving FOOL.



x.com/realstewpeters/status/1641848210330116096

Stew Peters
@realstewpeters

Morgan Freeman BLASTS Joe Biden for being an incompetent ice cream-loving FOOL.

TikTok
TikTokanofmanyvoices
Children are killed and this FOOL is talking about chocolate chip ice cream!!!!
0:03 / 0:41

Readers added context they thought people might want to know

Clearly not Morgan Freeman this is a celebrity impersonator aided with face filtering deep fake technology.

tiktok.com/@themanofmanyv...

Figure 2.6: Morgan Freeman criticizes President Biden (deepfake)
Source: https://x.com/realstewpeters/status/1641848210330116096

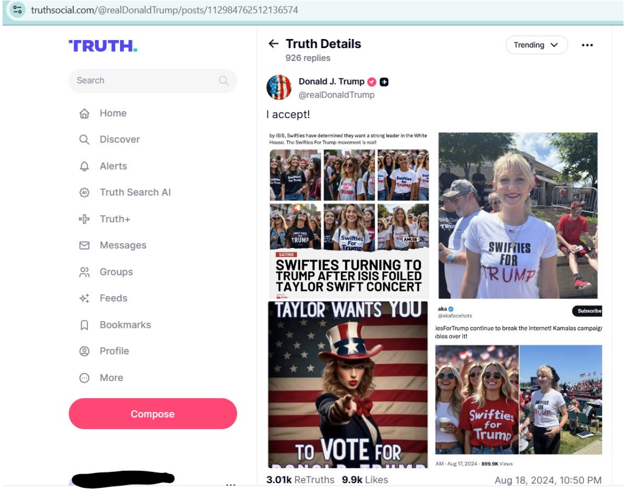
Once again, the author uses a well-known media technique to attract attention and evoke emotions: capital letters and dramatic words. The use of children in the text also evokes emotions, making use of a national tragedy to add another layer of criticism to the former President. This clearly has had an effect, as the post has gathered 5.3 million views and thousands of shares. Unsurprisingly, among the sharers, we find people expressing political partisanship, but also a lot of seemingly fake profiles. This time as well, however, many people also debunked the content. As for the poor video quality, the movements of Freeman appear very unnatural, as if a mask was superimposed on his face. What is more, if the video was genuine, one would expect it to be posted by its claimed author as well. However, the actor himself does not have a TikTok account.

This additional verification check is unlikely to be taken by most users, especially if they are emotional and are already prone to believing the suggested story about the President. In fact, as the idea of confirmation bias teaches us, people are more likely to accept the truth of news supporting their existing beliefs, while also discounting contradictory evidence. This becomes especially powerful on social media, where people often share headlines without reading them, relying on intuitive judgment rather than analytical thinking, especially when the content is emotionally charged or aligns with their views (Pennycook & Rand, 2019).

1.4 President Trump Endorsed by the Swifties

Another, more benign example can be observed on Truth Social – the social network of Donald Trump. It has been shared by Donald Trump himself in the context of his second electoral campaign. It is a compilation of screenshot posts from X users containing deepfake images of media articles and photos of young girls, seemingly fans of Taylor Swift, who demand a strong leader and are rallying against a Swifties for Trump movement. They also use capital letters in the caption and bait words such as “SHOCK”. The original posts have gained hundreds of thousands of views. Trump’s post is from August 2024, and it follows the cancellation of a Taylor Swift concert in Vienna due to possible terrorist attacks planned by ISIS. In reality, Taylor Swift had not endorsed Trump and had also criticized him publicly.

Trump has obviously combined a few posts and screenshots to steer public opinion in his favour, accompanying the post with a caption reading that he accepts being the strong leader in the White House. The post can contain some truth. Likely, the author of the original post is a Trump and a Swift fan, who has even crafted herself a t-shirt with the label Swifties for Trump. All other illustrative images, however, have obviously been generated with AI.



truthsocial.com/@realDonaldTrump/posts/112984762512136574

926 replies

Donald J. Trump
@realDonaldTrump

I accept!

by ISIS, Swifties have determined they want a strong leader in the White House. The Swifties For Trump movement is real!

SATIRE

SWIFTIES TURNING TO TRUMP AFTER ISIS FOILED TAYLOR SWIFT CONCERT

aka
@akafacehots

iesForTrump continue to break the internet! Kamalas campaig
ıbles over it!

AM · Aug 17, 2024 · 899.9K Views

[redacted]

3.01k ReTruths 9.9k Likes Aug 18, 2024, 10:50 PM

Figure 2.7: President Trump's post on Truth claiming endorsement from Taylor Swift's fans
Source: https://truthsocial.com/@realDonaldTrump/posts/112984762512136574.

The post has not accumulated that many views and shares, but it illustrated another possible use of deepfake content on social media: to fake support and endorsement of political candidates.

There is no evidence behind the intentions of Donald Trump, but there is solid scientific evidence showing that celebrity endorsements and influencer status can significantly increase the perceived credibility of fake news or misinformation, even when the content itself is misleading (Mena et al., 2020). A study using eye-tracking experiments demonstrated that articles featuring celebrity images and sensational headlines (fake news style) command more viewer attention than other content, even drawing attention away from the article's factual text. This signals a strong unconscious attraction to celebrity-linked fake content (Lazar & Pop, 2021). At the same time, it is known that celebrity amplification can cause real harm, something that has been documented multiple times during the Covid-19 pandemic, when influencers, including celebrities and wellness figures, played outsized roles in spreading anti-vaccination conspiracies, introducing personal narratives that increased engagement and made moderation more complex (Observatory, 2022).

1.5 Vladimir Putin talks to… Vladimir Putin

Finally, another example of a deepfake which has spread rapidly online, but this time for a different purpose: to educate the public, or rather, to convey a political narrative. The video comes from Russia and is state-sponsored. To address numerous rumours appearing in Western media, claiming that President Putin does not personally attend meetings, but uses doubles, the team behind the Russian president has decided to perform a media exercise and show how easy it is to be fooled by deepfakes. It shows real Vladimir Putin sitting in a studio with a live audience during his annual news conference. The president is looking at a screen, taking questions about policy from remote speakers. At some point, an AI-generated Putin lookalike appears, presenting himself as a student. He has the body and voice of Putin, and it therefore looks like the president is talking to his Doppelgänger. During the conversation, Putin's AI look-alike asks the president if he has a lot of doubles and his opinion about the dangers of deepfakes. The content originally appeared on national TV in December 2023 and only then spread to social media worldwide, making it impossible to track its exact spreading path. The numerous news headlines from large online media show that it made an impact. Alongside the purpose to inform and to spread fear that something happened to the Russian leader, this video also served the Russian-state propaganda goal of portraying Western media as biased against Russia, by using the very weapon Russia is usually blamed for using: disinformation.

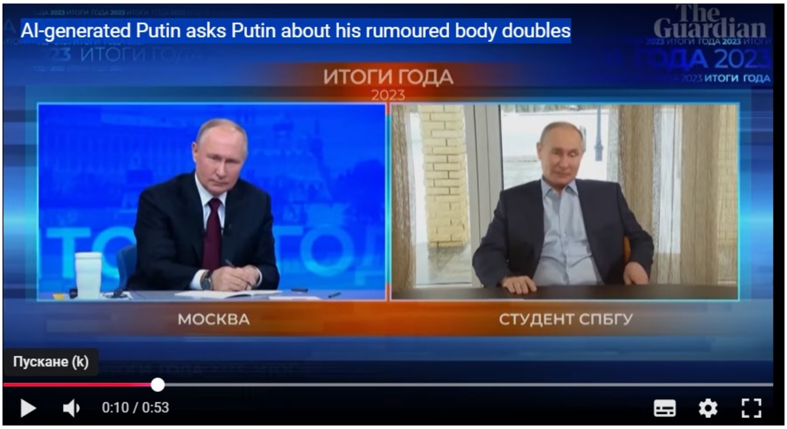


Figure 2.8: Putin talks to AI-generated Putin.
Source: The Kremlin via The Guardian
https://www.theguardian.com/technology/artificialintelligenceai/2023/dec/14/all.

1.6 Challenges

Most social media can become a vehicle of deepfake disinformation. Some of the key factors enabling this are the rapid content distribution, a trusted environment in closed groups, filter bubbles, echo chambers, anonymity, private chats, influencers, resulting in the empowerment of virtually all users to become media on their own.

Disinformation is an intentional act, with its authors usually choosing the best network depending on their needs. Engagement-driven algorithms of Facebook, for instance, keep showing us more of what we like, encouraging users to engage with similar content and causing stronger emotional reactions. Its large user base, which includes many users who are not used to detecting risk factors in digital environments, combined with current struggles to detect AI-generated disinformation and failure of automatic content moderation, makes Facebook an ideal ground for deepfakes disinformation. Platforms like X are doing better with flagging AI-generated disinformation and adding context, but the platform's dominant political and news orientation allows for politically motivated deepfakes to spread rapidly.

There are now many challenges to analysing how content spreads on social media to regular users or independent journalists. Previously easily accessible tools like CrowdTangle, a Facebook software allowing users to follow the spreading of online content, have been discontinued and replaced by less efficient and accessible alternatives (Gotfredsen & Dowling, 2024).[1] Notably, alternative tools for trend analysis and monitoring are available, but they are also expensive and usually require some degree of technical knowledge.

Most importantly, even if bots and fake profiles boost the distribution of a deepfake, a very concerning fact is that it is very often popular public figures, influential in the public space, who distribute deepfakes, exploiting emotions, patriotism, vulnerable groups, and sensitive social topics to serve their goals.

[1] Meta claims that Meta Content Library (MCL) is the new tool to provide high-quality data to researchers, while abiding by regulatory requirements for data sharing and transparency. However, reports claim that this tool is much less accessible, transparent and useful.

While social networks are incapable (or unwilling) to slow the spread of deepfakes, since their internal policies and one-size-fits-all interventions are proving too slow or inefficient, progress by experts promises to help tackle AI disinformation concerns. A step in this direction is represented by statistical approaches monitoring disinformation waves that, by identifying distinct, vulnerable populations, can then help to identify customized and more effective debunking interventions.

2 Statistical Approaches to Segmentation

The analysis of propagation dynamics and statistical detection models presented in this chapter provides the theoretical and technical framework necessary to interpret the case studies discussed in the previous section. While the latter examined the tangible effects of synthetic disinformation, such as the manipulation of public opinion through the falsified image of European leaders or the doctored video of President Biden, this section deconstructs the underlying mechanisms driving these phenomena. It becomes evident, for instance, that the virality of such content is not accidental, but rather the predictable result of the interplay between the engagement-driven algorithms described in the previous section and the emotional levers of fear or indignation that characterized those specific episodes.

Furthermore, the hybrid detection methodologies proposed in this section, grounded in sentiment analysis and time-series anomaly detection, directly address the critical vulnerabilities exposed in the previous examples. Where the human eye and traditional verification methods reached their limits against the visual hyper-realism of the Morgan Freeman deepfake or the rapid dissemination of falsehoods on Twitter, the statistical approach illustrated here offers a tool capable of identifying the latent traces of manipulation. Consequently, this section does not merely describe network operations: it proposes a methodological response to the systemic vulnerabilities exemplified by the narratives described previously.

As discussed in the previous section, the spread of synthetic media, especially deepfakes created with generative AI, has deeply changed the digital information landscape, creating serious challenges for truth, public debate, and democratic stability. What began as an innovative technology now enables the rapid and convincing spread of fake content, greatly strengthening disinformation efforts. Because of this, it is crucial to take a critical look at existing statistical methods, beginning with segmentation techniques that group people by their level of vulnerability, and moving toward advanced models that uncover the subtle social effects of AI-generated false information.

At the core of these developments is the need to view the digital information space as a complex ecosystem shaped by many different actors. These actors include individual users, each with distinct cognitive styles, emotional traits, and levels of trust in media, as well as collective agents such as social media platforms, algorithms, automated bots, and influential content creators. Together, they shape the speed, scale, and spread of synthetic media, including deepfakes and other forms of advanced misinformation.

This complexity creates the need for a comprehensive analytical framework integrating micro-level processes, such as individual susceptibilities, cognitive biases, and emotional responses, with macro-level systemic structures like networks and algorithmic affordances. Only by jointly examining these dimensions can researchers map how vulnerabilities emerge, disseminate, and embed in the digital milieu.

Advanced statistical modelling plays a key role in examining the diversity and variation within a population. The psychological foundations discussed in Chapter 4 will later explain how sociodemographic, motivational, and cognitive factors shape people's susceptibility to deepfakes techniques such as logistic regression, latent class analysis (LCA), factor analysis, clustering algorithms (used to group similar things together), and structural equation modelling (SEM) allow researchers to extract latent psychological and behavioural profiles from complex datasets. These tools identify distinct risk groups and reveal how interconnected beliefs, emotions, ideologies, and digital engagement cultivate susceptibility (Bhatnagar & Ghose, 2004; Kang et al., 2020; Outwater et al., 2003; Verma, 2013; Yan et al., 2018).[2]

However, current models have some limits. They often look at only a few factors and rely too much on data from Western countries. To make them more useful, researchers need to include data from more regions and cultures and use long-term, cross-platform studies to track how people's vulnerability changes over time.

[2] Logistic regression statistical method that predicts the probability of something happening and turns that into a yes/no decision. LCA is used to find hidden groups (or "classes") within a set of people (or items) based on their answers, behaviours, or characteristics. Factor analysis is used to find underlying patterns or "factors" in a large set of variables. It helps researchers understand which variables go together and what hidden dimensions explain them. Finally, SEM is a powerful statistical method used test complex cause-and-effect relationships between observed and hidden (latent) variables, all at once, in a single, comprehensive model.

Therefore, building an effective and lasting response to AI-driven disinformation requires collaboration across different fields, combining insights from psychology, statistics, computer science, and socio-political studies. This well-rounded approach is crucial for identifying where people are most vulnerable and developing evidence-based strategies that strengthen democratic resilience in a constantly changing information environment.

2.1 Statistical Modelling Approaches for Studying the Impact of GenAI Content and Fake News

Recent advances in statistical modelling have substantially deepened our understanding of the multifaceted and often subtle ways in which AI-driven synthetic misinformation spreads, affects, and reshapes different societal groups. Researchers now use a wide range of sophisticated quantitative methods to uncover the multiple, context-dependent factors that drive susceptibility, moving beyond basic descriptive analyses toward detailed modelling of influence networks, belief formation, and behavioural dynamics (Sæbø et al., 2020).

Together, these statistical methodologies unlock unprecedented insights into the complex factors driving the spread and societal impact of AI-generated fake news. They allow researchers to map intricate networks of influence, which are often shaped by automated bots, coordinated influencer campaigns, and opaque platform algorithms, and translate this knowledge into practical, evidence-based solutions. These solutions range from carefully targeted media literacy programs designed for specific risk groups to predictive tools that identify emerging vulnerability clusters, to real-time content detection and moderation systems that can interrupt misinformation cascades at critical points, as well as adaptive regulatory measures that help platforms and policymakers respond quickly and effectively to the evolving disinformation landscape.

The true power of statistical tools lies in their ability to integrate theory and practice: turning conceptual understanding into evidence-based, context-sensitive interventions that help civil society and institutional actors detect, anticipate, and counter the harms caused by synthetic media. In an era defined by the rapid evolution of generative AI and the growing sophistication of synthetic content, only a continuously adaptive, data-driven, and theoretically grounded approach can protect the integrity of knowledge and strengthen democratic resilience in digital public spaces, thereby safeguarding the foundations of informed citizenship in the twenty-first century.

At the forefront of this endeavour stands logistic regression, a versatile statistical tool pivotal in isolating and quantifying individual-level risk factors (Shete et al., 2021). Variables such as age, educational background, ideological leanings, and media consumption patterns are no longer treated as mere demographic markers but are examined as dynamic mediators and moderators situated within complex psychosocial ecosystems. For instance, the protective influence of education may depend heavily on a person's digital literacy, while political ideology can influence news consumption and openness to misinformation in complex, non-linear ways. By incorporating these factors within interacting cognitive and sociocultural networks, logistic regression provides a nuanced understanding of how vulnerabilities emerge, showing how individual predispositions interact with structural exposures to increase susceptibility to fake news.

Latent class analysis (LCA) expands analytical possibilities by moving beyond predefined groups to reveal hidden subpopulations whose vulnerabilities stem from unique combinations of beliefs, emotional traits, and media engagement patterns (Shen & Wu, 2024).

This method is particularly effective at revealing the fluid and overlapping nature of audience segments that cannot be easily captured by simple demographic or psychographic categories. For example, LCA can identify clusters of users whose exposure to synthetic media is shaped by the combined effects of cultural norms, peer influence, and algorithmically curated content, together creating hidden vulnerability profiles. This approach reframes susceptibility not as a fixed individual trait but as a dynamic interaction of self-concept, social identity, technological mediation, and the broader networked environment, highlighting the need for innovative segmentation models and precisely targeted interventions.

Adding another layer of methodological sophistication, structural equation modelling (SEM) allows researchers to estimate both direct and indirect causal pathways connecting a complex set of cognitive, emotional, and socio-structural variables (Tahat et al., 2022). SEM is particularly effective at analysing the recursive and often bidirectional feedback loops found in digital misinformation ecosystems. It maps complex relationships, such as how media trust directly influences credulity, or how ideological alignment affects the emotional impact of deceptive content. For example, SEM can model how initial acceptance of a deepfake sparks emotional arousal, which then increases selective sharing and fosters attitudinal polarization within networked communities. This level of analytical detail is essential for understanding the self-reinforcing dynamics that drive the spread and lasting impact of synthetic media among digitally connected audiences.

2.2 Case Study: Early Detection of Fake News through a Hybrid Statistical Framework

Within the SOLARIS project, we developed an innovative hybrid statistical model designed to enhance the identification of AI-generated fake news. This approach integrates diverse analytical techniques to improve both the accuracy and timeliness of detecting synthetic misinformation within dynamic digital environments.

Our methodology operates on two complementary levels. The first one focuses on analysing the emotional tone of news articles using sentiment analysis (Mohammad & Turney, 2013). Here, we measure the expression of key emotions such as fear, anger, sadness, and trust throughout a text. It is consistently observed that fabricated news exploits emotional manipulation, often intensifying negative emotions like fear and anger to capture reader attention and influence perceptions. By assessing patterns of emotional intensity and variability, we distinguish characteristic differences between real and fake news; as suggested in the previous section, authentic journalism generally maintains a balanced and steady emotional tone, whereas misinformation reveals abrupt spikes in distressing sentiments.

The second level concentrates on behavioural data, specifically analysing public engagement through online search trends. For instance, we monitored monthly search interest for the term “nuclear” spanning from 2004 to 2025 (see Figure 2.9 below). Sudden, anomalous surges in search volumes signal potential misinformation events or coordinated disinformation campaigns igniting public concern.

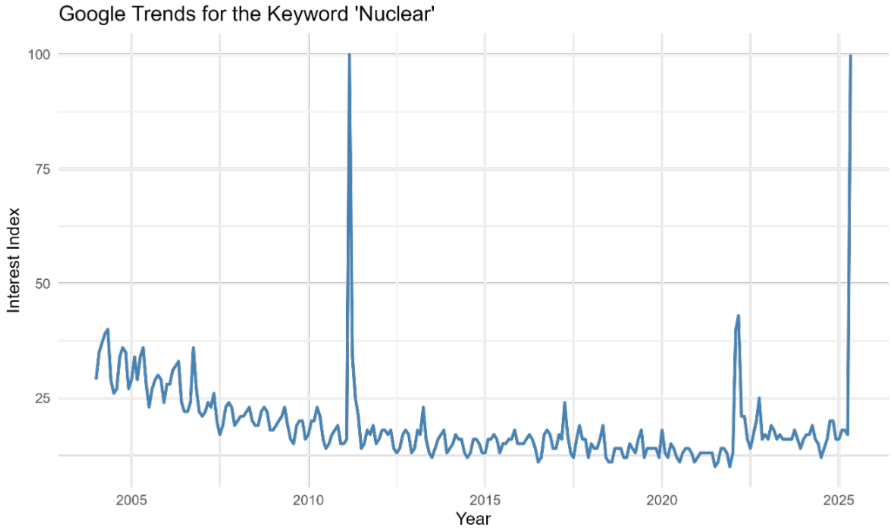


Figure 2.9: Monthly Google Trends data for the keyword nuclear (2004–2025). The final observation is artificially adjusted to simulate an anomalous spike.
Source: Fenga and Biazzo, 2025.

To robustly detect such anomalies, we deploy multiple forecasting models, including traditional time series techniques, such as Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ETS), alongside advanced machine learning models like the Extreme Learning Machine (ELM) neural network (Chatfield et al., 2001; Shumway & Stoffer, 2017; Wang et al., 2022).[3] We further enhance reliability using bootstrap resampling methods to generate confidence intervals, defining expected "safe zones" of variation against which real-time observations are evaluated (Hesterberg, 2011).[4] Once observed search frequencies exceed these bounds, the system flags a possible fake news event.

In experimental evaluations, we constructed a dataset comprising 20 genuine news articles alongside 5 AI-generated fake news pieces, paired with corresponding Google Trends data. Artificially injecting anomalous spikes into the search data, we tested the system's detection efficacy. The sentiment analysis reliably separated

[3] ARIMA is used to predict future values in a time series – like stock prices, weather, or website traffic – based on past data. ATS is a method for forecasting future values in a time series by giving more weight to recent observations and less weight to older ones. Finally, ELM is a type of artificial neural network used for classification or regression tasks – basically, for predicting outcomes or categorizing data

[4] Bootstrap resampling allows researchers to estimate the reliability of a statistic by repeatedly sampling from data, even if they do not know the underlying population.

fabricated from authentic content, evidencing higher levels of negative emotion and volatility in fake news. Concurrently, all forecasting models successfully and synchronously detected the synthetic anomaly, without false alarms during baseline periods, confirming the system's sensitivity and robustness.

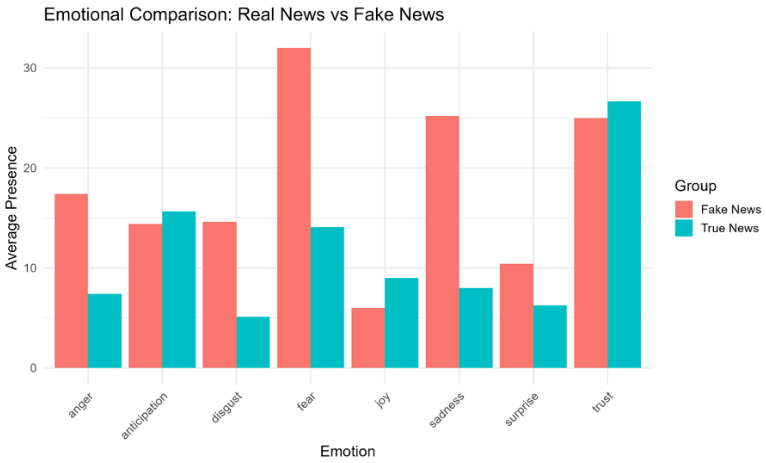


Figure 2.10: Relative emotion activation frequencies. Fake news intensifies fear, anger, and sadness.
Source: Fenga and Biazzo, 2025.

This dual-layered framework offers a potent early-warning tool against the proliferation of fake news. By uniting semantic emotional insights with behavioural metrics derived from real-time search activity, the model facilitates timely alerts for journalists, fact-checkers, and digital moderators, allowing for swift responses to emerging disinformation. Importantly, it is conceived as an augmentation rather than a replacement of human expertise, providing prioritized signals that guide investigative and corrective action. Its modular design permits adaptation across diverse languages and topical domains, enhancing its versatility and broad applicability.

2.3 Future Directions

Looking forward, combining advances in theory, statistics, and computation creates a strong research agenda to address AI-driven synthetic misinformation. As generative technologies increasingly blur the line between reality and fabrication, current models show important limitations and highlight the need for interdisciplinary approaches. Understanding vulnerabilities will require integrating psychological, behavioural, technological, and socio-political factors, as well as conducting long-term and cross-cultural studies. Real-time analytics and advanced natural language processing can support predictive and responsive interventions, helping policymakers and platforms act quickly when misinformation threatens social cohesion and democracy. At the same time, robust ethical frameworks and regulations are essential to protect privacy, rights, and public trust amid widespread digital manipulation. By building an adaptable, integrated framework that combines diverse data sources and methods, we can strengthen societal resilience against fake news and safeguard the integrity of public discourse and democratic institutions in this fast-changing digital era.

3 Detecting deepfakes on social media: the perspective of journalists and press agencies

For journalists and especially for freelancers, who often work alone with limited resources and under tight deadlines, the rise of deepfakes represents one of the most daunting and complex challenges faced in recent years; the same years in which an unprecedented technological revolution has profoundly transformed the world of information and, with it, the way the public reads and understands the present (Sohrawardi et al., 2020).

First came the pervasive spread of social networks such as X (formerly Twitter), Facebook, or Reddit: platforms whose algorithms decide what we see and when, based on criteria that are anything but transparent. These platforms have radically changed the way news is consumed, polarising opinions and systematically promoted "viral" content that generates engagement and, with it, valuable data for the very companies that produce and monetize these social networks. At the same time, the success of instant messaging systems such as WhatsApp, Telegram, or (to a lesser extent) Viber and Signal has created new spaces for exchange and sharing, such as channels and groups, where all kinds of content, including deepfakes, can be shared and reshared virtually without control (Al-Khazraji et al., 2023).

Now, adding to this landscape already extremely complex for journalists to decode, comes the unstoppable and rapid evolution of AI tools capable of generating fake audio and, above all, video content that is increasingly realistic, carefully crafted to go viral. It is a perfect storm putting great strain on a profession built on testimony and fact-checking.

The risk for journalists, and especially for freelancers, is twofold: on the one hand, there is the danger of falling into the trap after receiving an apparently authentic and relevant video, audio clip, or image (such as a fragment of a private conversation between politicians or an inconvenient admission by a public figure) and relaying it, thus becoming an unwitting cog in the disinformation machine. The urgent need to "stay on the story" and be the first to publish represents a shared necessity for both freelance and editorial journalists, with the major difference being the absence of a structured editorial team for cross-checking information for the former. A difference that can play a decisive role in the fight against disinformation and hinder professional integrity. The result: reputational damage that, for an individual professional, can be irreparable.

On the other hand, there is a subtler but equally insidious challenge: hyper-scepticism. When everything can be fake, verification work turns into an exhausting investigation. While the pillars of journalism, such as cross-checking authoritative sources or analysing context, remain the foundation of reporting, when every audio or video file becomes suspect, verification requires a process that drastically slows down the workflow, all while the "news" spreads uncontrollably across social networks. It is no longer just about cross-checking sources or verifying a witness's credibility, and about analysing a file's metadata and hunting for micro-imperfections in a video, such as an unnatural blink, a strange blur along the edge of a face, or inconsistent lighting. These details are becoming increasingly difficult to spot due to the progress of generative AI, as shown for instance, by the recent release of Veo 3, Google's video generator based on Gemini AI, which has "broken the silence barrier" by adding audio to ever-higher-quality images.

Fortunately, the same AI that creates the problem also provides part of the solution. Today's freelance journalist must necessarily combine a nose for news with technological competence. There exist AI tools specifically designed to detect deepfakes, and information professionals must learn to use them just as they once did with a notebook. Platforms such as Reality Defender, free software such as Deepfake-O-Meter, IdentifAI, or Sentinel (more suitable for companies and institutions), for example, analyse multimedia files submitted to them in search of digital artifacts and inconsistencies invisible to the human eye (Stephen, 2025).

Others focus on discrepancies between mouth movements (visemes) and spoken sounds (phonemes), a detail almost impossible to counterfeit perfectly.

However, the possibility of escalating (allegedly) fake news to other members of the editorial team points to the fact that, surprisingly enough, technology represents a last resort. These tools represent valuable support, but they cannot replace (and probably never will) human judgment and established journalistic practices: editorial journalists themselves tend to first cross-check with other sources reporting on the news, leveraging their newspaper's connections. By leveraging contacts with other newspapers, press offices, spokespersons, institutional social media profiles, and so on, editorial journalists are able to determine whether events depicted through a deepfake actually took place in the real world.

Second, journalists look at the context of the news. For instance, in case a public figure (such as a politician) were to give a speech that does not resonate with their known stance on the topic, say, a climate change denial message from activist Greta Thunberg, journalists may already flag the news as suspicious and, once again, check with other sources.

Finally, technical features of the video may be highlighted as suspicious by the expert eye of journalists, who may, for instance, detect discrepancies between mouth movements and spoken sounds, details almost impossible to convincingly counterfeit. Only at this point may editorial journalists resort to the help of detection software to analyse media content and determine whether the video depicts real or made-up events. This is the case, for instance, when journalists cover war areas, where it is difficult to cross-check with other sources or to extrapolate enough information from the context where events unfold.

In contrast to editorial journalists and the resources available to them, freelance journalists are able to resort to a multi-level approach: technology for initial screening, followed by a critical contextual analysis that only a journalist with the right expertise can provide. The fundamental question for both editorial and freelance journalists, however, remains the same: *cui prodest?* Who benefits from the spread of that false content? Then, as always, the process continues by cross-checking the news with known facts, testimonies, and primary sources. In short, navigating this constantly evolving landscape requires a new form of "augmented journalism": freelancers (as well as newsroom journalists) must become more meticulous, more transparent in their verification process, and above all, humbler. They must be ready to admit what cannot be verified with certainty and to explain to their audience the complexities of an information ecosystem where distinguishing between truth and falsehood has become the new, crucial challenge to overcome.

3.1 Use Case 2 – The SOLARIS Project Disinformation Event

Pursuing the goal of empowering journalists with relevant tools and skills to combat AI disinformation, the SOLARIS consortium organized a brainstorming session at ANSA's headquarters in Rome, involving journalists, communication experts, institutional representatives, researchers, private companies" professionals, and different stakeholders from the information sector. More specifically, the objectives of the event were as follows:

- collect feedback on how ANSA journalists detect and manage deepfakes in their daily work,
- co-design mitigation strategies, and
- formulate concrete recommendations to address "infodemics."

During the two days in which the roundtable debate took place, participants attended an editorial meeting to closely observe the daily working process of ANSA journalists of the newspaper agency's key activities. This allowed to witness the established processes and criteria by which ANSA decides which stories to cover and how to develop their reporting. Following the editorial meeting, a group of senior ANSA journalists was shown three deepfakes created specifically for the event: the goal was to assess their reactions and response procedures, as well as to identify possible gaps in current practices.

The debate then expanded into a session involving experts and the different kinds of stakeholders mentioned above, who started by identifying different types of AI-generated disinformation and their varying implications. Subsequently, the working group turned to the search for solutions, reflecting on the role of human beings in using their professional experience to combat disinformation and on the possibility of fighting fire with fire – that is, using AI to detect fake news, to promote digital literacy, and to create counter-narratives against deepfakes disinformation.

The event concluded with an interactive session in which ANSA journalists further discussed with experts the potential of detection tools and the adequacy of current laws and regulations targeting online disinformation.

3.2 Traditional Journalism vs. Deepfakes

The good news, then, is that professional journalism (especially with the support of the resources and practices of editorial settings), with its layered processes and models, already has many effective tools to counter deepfakes. The SOLARIS roundtable, in fact, highlighted a multi-level verification approach to identify and neutralize any false or manipulated content, including deepfakes. This process does not rely on a single tool, but rather on a combination of technical analysis, in-depth contextual knowledge, and rigorous journalistic principles.

The initial analysis of suspicious content often starts with superficial warning signs, such as evident imperfections in terms of context (missing or incorrect source logos), content (such as, for instance, a politician expressing a political stance incoherent with their long-held political beliefs), or obvious technical errors, like poor synchronization between audio and video. However, participants present at the brainstorming session stressed that the technical quality of a video is neither the only nor the most important evaluation factor: eventually, the true core of their defence strategy is keeping the human component at the forefront of technology use to tackle disinformation: journalistic experience makes the difference. Deep knowledge of specific contexts, sources, and public figures generally enables journalists to detect anomalies that an algorithm or an inexperienced eye would not be able to catch.

The network of regional correspondents and collaborations with other international news agencies (such as the BBC) acts as a cross-checking mechanism, essential for validating doubtful information, although editorial journalists argued they would not have cross-checked with other critical sources to verify the news, since technical, content, and contextual details all strongly pointed to the made-up nature of the videos analysed. Ultimately, ANSA journalists argued that the strongest defence lies in the core principles of journalistic work. Editors reiterated that source attribution is a fundamental and non-negotiable requirement. In an era of viral disinformation, the newsroom deliberately chooses to prioritise accuracy over speed, a principle that translates into the need to verify every story through direct contact with sources and to always seek multiple confirmations before publication.

Emerging from SOLARIS discussions, the key steps journalists may take against deepfakes can be summarized in the following order:

- The ability to cross-check online information with other media outlets or relevant institutions is at the heart of debunking disinformation.
- The content of deepfakes may provide very important hints: if the content is plausible, journalists need to leverage on their expertise to verify whether there exist inconsistencies in the message conveyed through the video.
- The context in which a video is set also delivers key insights about the content's credibility. With context, technical details (and journalists' ability to recognize them) become critical to detect fake news. Additionally, war contexts make videos more difficult to cross-check.
- Finally, supporting experts to identify technical inconsistencies, detection technologies may complement traditional processes with modern verification tools, including detection software based on AI.

The SOLARIS event also underlined the importance of a clearer taxonomy of disinformation. The discussions highlighted the crucial importance of distinguishing between "disinformation" and "AI-generated disinformation" – the latter encompassing video, audio, or written sources at an output-intensive pace compared to traditional disinformation – and "misinformation," the unintentional sharing of what is believed to be true, as well as "malinformation," which amplifies disinformation with defamatory intent. From the debate it emerged the need to differentiate "harmful content" according to its degree of risk.

Finally, among the critical issues that emerged from the dialogue between journalists and experts was also a worrying decline in public trust towards traditional media. To rebuild this trust – the panel suggested – it is essential to actively involve citizens rather than imposing knowledge from above. This can also be achieved by focusing on coaching professionals and end-users to understand the positive impact of generative AI on disinformation, which aims to use AI to detect deepfakes and generate content to develop counter-narratives to false news. More broadly, media literacy campaigns were recognized as a crucial tool to restore public trust and prepare citizens to navigate an increasingly complex information landscape.

4 Mitigating: Slowing the Spread

In the recent generative AI (genAI) wake, social scientists have pointed to the skill-replacing threat of AI technology over its skill-enhancing potential: people's ability to develop essential skills such as critical reading and structured thinking is hindered by the possibility to delegate tasks to AI tools, which makes education-related efforts appear redundant. Among other things, this translates to individuals being ill-equipped with the necessary knowledge to identify and react to online disinformation (Arribas et al., 2025). The affirmation of deepfakes as increasingly trustworthy visual content magnifies disinformation risks related to human-artifact interaction in the online context.

Citizens' inability to learn about and defend themselves from deepfakes hinders their status as rights-holders, eroding their capacity to self-advocate for the principles of transparency, privacy, and accountability. At the same time, deepfakes risk weakening democratic participation, widening social gaps by increasing the digital divide (Lythreatis et al., 2022). Against the backdrop of AI as a vector of technological disruption, experts have stressed the importance of democratising the values behind the introduction of AI tools: if citizens are to benefit from social media platforms and AI tools as a means for enhancing democratic engagement in the online context by combating disinformation, better inclusion of most diverse categories of citizens is most desirable in order to help identify socially critical AI problems (Corrêa & Oliveira, 2021).

However, the bottom-up approach must also be matched by efforts at empowering citizens with relevant knowledge on AI and deepfakes. By stressing the peculiarities of AI as a fast-changing technology, the limits of top-down regulatory approaches and institutional initiatives, the role of AI education as a precondition for enhancing the fight against AI-generated disinformation and strengthening individual rights in the online context is advocated for.

Economides (1996) and Birke (2009), focusing on Information and Communication Technologies, show that as more people adopt a network technology, its performance improves (Birke, 2009; Economides, 1996). AI systems exhibit this network externality too: the larger the data network they access, the more intense their training (LeCun et al., 2015; Panno et al., 2023). Learning-oriented algorithms nonetheless tend to go beyond what network technologies traditionally envisage in terms of spillover effects: in this case, the network features dramatically increase AI's ability to autonomously enhance its output (Levine & Jain, 2023). This, of course, also improves deepfakes' ability to mislead. The possibility to quickly create increasingly trustworthy deepfakes interacts with the global reach of world-famous platforms, such as those owned by Meta, which have occasionally contributed to political misinformation and disinformation dynamics (Acemoglu et al., 2025).

These problems have been approached by tightening the regulatory stance of national institutions. The EU context is usually taken as a benchmark comparison, considering the proactive regulatory stance the 27 have taken to address these problems. Legislative projects such as the AI Act and the Digital Services Act (DSA) have focused on preventing the introduction of AI technology deemed dangerous for end-users and on extending accountability of online platforms in terms of illegal and harmful content that may circulate through their digital environments. These initiatives mostly focus on engaging with technology producers, setting normative standards for the production of safe AI services. Alongside binding documents, the EU has also attempted to encourage voluntary compliance to safe information standards through the 2022 Strengthened Code of Conduct on Disinformation – integrated in the DSA in 2025 (European Commission, 2025). Such legal documents, however, do not yet appropriately tackle laypeople's AI education and critical skill development. Communication experts and journalists are therefore left to bridge the AI-generated information gap by either flagging fake content or by fact-checking the content of deepfakes (Painter, 2023). Forja-Pena et al. (2024) nonetheless stressed how newspapers are currently navigating the challenges posed to their working category from AI, investigating the ethical and efficient use of AI technologies to contrast disinformation and to help produce quality information (Forja-Pena et al., 2024). At the same time, they also highlight the lack of adequate technological literacy to tackle online disinformation and assist journalists in their jobs of quality reporting. Nonetheless, they also highlight the lack of adequate technological literacy to tackle online disinformation and assist journalists in their jobs of quality reporting. This represents a notable shortcoming in the fight against online misinformation, disinformation, and malinformation.

Even though AI education represents an urgent goal to be pursued in the context of combating disinformation, the delay in dissemination programmes stems from the ongoing debate on what constitutes relevant AI knowledge (Hermann, 2022; Kandlhofer & Steinbauer, 2018; Long & Magerko, 2020; Mikalef & Gupta, 2021):

what are the necessary notions to navigate a rapidly changing, self-enhancing technology? Given the dynamic nature of AI, would a theoretical and general preparation represent a better option than practical, AI tool-specific knowledge?

In the attempt to identify helpful AI notions, there exist governmental initiatives that have promised to prepare civil society to engage with AI tools and to promote political participation and the upholding of democratic values for digital citizens. By collecting citizens' input, such initiatives aim to inform the government's ability to support and provide adequate education and solve context-dependent problems of GenAI applications. A relevant instance of this political experiment comes from the Kingdom of the Netherlands, where the "Government-wide vision on generative AI of the Netherlands" advocates for country-wide resilience to AI-related challenges (Zaken, 2024). The resort to civil debate initiatives, such as the AI Parade, aims to collect data from citizens' experiences with AI technology, to articulate the goals of an AI education whose necessary knowledge is framed directly by digital citizens' needs.

Although the Dutch initiative does not revolve around the specific topic of AI disinformation, the constructivist approach of societal dialogue represents an important attempt at closing the information gap, at pursuing timely AI education, and at safeguarding democratic functions and norms. Providing citizens with the opportunity to share hands-on AI knowledge and to voice the expectations with respect to the introduction of different kinds of AI products and services is an unavoidable step, and it has been recognized as such by international stakeholders, even if this dialogue has mainly been understood from the perspective of preventing a worsening of the working conditions in relation to the introduction of AI (Cazes, 2023; Krämer & Cazes, 2022). Still, better regulation from institutions and enhanced cooperation by social media platforms are understood as the necessary and sufficient condition, or to the very least as the most urgent measure, to protect digital citizens and democratic institutions, with no complementary role envisaged for societal dialogue, AI education, and knowledge-sharing on online experiences (Painter, 2023; Pawelec, 2022).

Nonetheless, AI knowledge sharing is pivotal to the debate on a human-centred AI – that is, an ethical introduction of AI tools that enhance human capabilities rather than substituting them – and to the current regulatory focus behind strengthening democratic values and fostering ethical technological innovation (Khutsishvili, 2024). Therefore, the pursuit of civil debate and of knowledge sharing represents not a complementary and necessary component of tackling AI-generated disinformation, but an intrinsic element to the regulatory efforts and the scientific debate surrounding GenAI. Promoting a bottom-up AI education allows to tackle the legislative gap, to enhance efforts by journalists and fact-checking institutions, and to empower digital citizens to defend their rights.

5 Concluding Remarks

Deepfakes spread rapidly on social media by exploiting emotional responses, platform algorithms, and the authority of influential figures. The case studies examined illustrate how synthetic media can distort political discourse, cultural narratives, and public trust, often leaving lasting impressions even after exposure is corrected.

The statistical models and hybrid detection frameworks developed under SOLARIS represent innovative dual-layer research tools that merge computational linguistics with predictive analytics to detect disinformation patterns in real time. Specifically, our framework integrates sentiment analysis algorithms, which map emotional signals in text and identify manipulative spikes in fear, anger, or distrust, with advanced statistical forecasting models such as ARIMA, Exponential Smoothing (ETS), and machine learning techniques that track abnormal patterns in public engagement data. For example, if reports of an alleged "nuclear incident" emerged, the system would simultaneously analyse the emotional tone of the content against established thresholds while monitoring surges in Google search activity that exceed statistical confidence limits. These combined signals generate quantitative alerts, allowing experts to prioritise potentially fabricated content before it spreads widely. In doing so, this approach shifts disinformation detection from reactive fact-checking to proactive monitoring, functioning as a comprehensive "statistical radar" that unites textual manipulation analysis with audience behaviour across multiple languages and topics.

While statistical models and hybrid detection frameworks offer promising tools for identifying vulnerabilities and anomalous patterns, they remain limited by technological, cultural, and methodological constraints. Journalists, particularly freelancers, face a dual challenge: avoiding uncritical amplification of deepfakes while also resisting hyper-scepticism that undermines timely reporting. Evidence from SOLARIS activities underscores the enduring importance of human expertise, contextual knowledge, and professional standards as safeguards against manipulation. Effective mitigation requires an integrated strategy combining advanced detection tools, enhanced media literacy, regulatory frameworks, and stronger accountability mechanisms for platforms. Persistent obstacles such as filter bubbles, opaque algorithms, and declining trust in traditional journalism complicate these efforts. To tackle such challenges and safeguard democratic discourse in the digital age, empowering citizens to critically engage with digital content involves yet another key stakeholder in the fight against disinformation.

End notes

Aseniya Dimitrova led the overall structure, writing, and editing of the chapter, and prepared the Introduction, the How Deepfakes Go Viral, and the Conclusion sections. Tommaso Tonello provided detailed feedback and contributed to editing the chapter. He authored the section on Mitigating the Spread of Deepfakes on Social Media, co-authored other parts of the text, and supervised reference formatting for the entire chapter. Luca Biazzo and Livio Fenga jointly conceptualized the statistical framework and authored the sections on Statistical Approaches to Segmentation, Advanced Statistical Modelling Methodologies, and Future Research Directions. They also designed and developed the innovative model for early detection of AI-generated fake news presented in the case study. Alessio Jacona contributed the perspectives of journalists and press agencies and summarized the findings from the SOLARIS Use Case in Rome. All authors reviewed and approved the final version of the chapter.

References

Acemoglu, D., Ozdaglar, A., & Siderius, J. (2025). *AI and social media: A political economy perspective* (No. w33892). National Bureau of Economic Research. https://www.nber.org/papers/w33892

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.

Al-Khazraji, S. H., Saleh, H. H., Khalid, A. I., & Mishkhal, I. A. (2023). Impact of deepfake technology on social media: Detection, misinformation and societal implications. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics, 23*, 429–441.

Arribas, C. M., Arcos, R., & Gertrudix, M. (2025). Rethinking education and training to counter AI-enhanced disinformation and information manipulations in Europe: A Delphi study. *Cogent Social Sciences, 11*(1), 2501759. https://doi.org/10.1080/23311886.2025.2501759

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

Birke, D. (2009). The economics of networks: A survey of the empirical literature. *Journal of Economic Surveys, 23*(4), 762–793. https://doi.org/10.1111/j.1467-6419.2009.00578.x

Cazes, S. (2023). Social dialogue and collective bargaining in the age of artificial intelligence. *OECD Employment Outlook*, 221.

Chatfield, C., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2001). A new look at models for exponential smoothing. *Journal of the Royal Statistical Society: Series D (The Statistician), 50*(2), 147–159. https://doi.org/10.1111/1467-9884.00267

Corrêa, N. K., & Oliveira, N. de. (2021). Good AI for the present of humanity: Democratizing AI governance. *AI Ethics Journal, 2*(2). https://doi.org/10.47289/AIEJ20210716-2

Economides, N. (1996). The economics of networks. *International Journal of Industrial Organization, 14*(6), 673–699.

Fenga, L., & Biazzo, L. (2025). *A hybrid statistical framework for early detection of fake news.*

Forja-Pena, T., García-Orosa, B., & López-García, X. (2024). A shift amid the transition: Towards smarter, more resilient digital journalism in the age of AI and disinformation. *Social Sciences, 13*(8), 403.

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

Hermann, E. (2022). Artificial intelligence and mass personalization of communication content – An ethical and literacy perspective. *New Media & Society, 24*(5), 1258–1277. https://doi.org/10.1177/14614448211022702

Hesterberg, T. (2011). Bootstrap. *WIREs Computational Statistics, 3*(6), 497–526. https://doi.org/10.1002/wics.182

Kandlhofer, M., & Steinbauer, G. (2018). A driving license for intelligent systems. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1). https://ojs.aaai.org/index.php/AAAI/article/view/11399

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

Khutsishvili, K. (2024). From a smart city to wise citizens: Smart empowering with artificial intelligence. In *Smart cities to smart societies* (pp. 51–63).

Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003439325-5

Krämer, C., & Cazes, S. (2022). *Shaping the transition: Artificial intelligence and social dialogue* (OECD Social, Employment, and Migration Working Papers No. 279).

Lazar, L., & Pop, M.-I. (2021). Impact of celebrity endorsement and breaking news effect on the attention of consumers. *Studia Universitatis, Vasile Goldis, Arad – Economics Series, 31*(3), 60–74. https://doi.org/10.2478/sues-2021-0014

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Levine, S. S., & Jain, D. (2023). How network effects make AI smarter. *SSRN Electronic Journal.*https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5281829

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

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

Lythreatis, S., Singh, S. K., & El-Kassar, A.-N. (2022). The digital divide: A review and future research agenda. *Technological Forecasting and Social Change, 175*, 121359.

Mena, P., Barbe, D., & Chan-Olmsted, S. (2020). Misinformation on Instagram: The impact of trusted endorsements on message credibility. *Social Media + Society, 6*(2), 2056305120935102. https://doi.org/10.1177/2056305120935102

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management, 58*(3), 103434.

Mohammad, S. M., & Turney, P. D. (2013). *NRC emotion lexicon.* National Research Council Canada.

Observatory, S. I. (2022). *Memes, magnets and microchips: Narrative dynamics around COVID-19 vaccines.* Virality Project.

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

Painter, R. W. (2023). Deepfake 2024: Will *Citizens United* and artificial intelligence together destroy representative democracy? *Journal of National Security Law & Policy, 14*, 121.

Panno, A., Pellegrini, V., De Cristofaro, V., & Donati, M. A. (2023). A measure of positive and negative perception of migration: Development and psychometric properties of the Positive and Negative Perception of Immigrants Scale (PANPIS). *Analyses of Social Issues and Public Policy, 23*(1), 73–105. https://doi.org/10.1111/asap.12338

Pawelec, M. (2022). Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital Society, 1*(2), 19.

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.

Sæbø, H. V., Ragnarsøn, R., & Westvold, T. (2020). Official statistics as a safeguard against fake news. *Statistical Journal of the IAOS, 36*(2), 435–442. https://doi.org/10.3233/SJI-190563

Shen, X.-L., & Wu, Y. (2024). Multidimensional information literacy and fact-checking behavior: A person-centered approach using latent profile analysis. In I. Sserwanga et al. (Eds.), *Wisdom, well-being, win-win* (Vol. 14597, pp. 280–297). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57860-1_20

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/

Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. In *Time series analysis and its applications* (pp. 75–163). Springer. https://doi.org/10.1007/978-3-319-52452-8_3

Sohrawardi, S. J., Seng, S., Chintha, A., Thai, B., Hickerson, A., Ptucha, R., & Wright, M. (2020). Defaking deepfakes: Understanding journalists' needs for deepfake detection. *Proceedings of the Computation + Journalism 2020 Conference*, 21. https://www.usenix.org/system/files/soups2020_poster_sohrawardi.pdf

Stephen, G. (2025). *Investigation and prevention of cybercrimes using artificial intelligence*.https://www.theseus.fi/handle/10024/891045

Tahat, K., Mansoori, A., Tahat, D. N., Habes, M., Alfaisal, R., Khadragy, S., & Salloum, S. A. (2022). Detecting fake news during the COVID-19 pandemic: A SEM-ML approach. *Computers, Integrated Manufacturing Systems, 28*(12), 1554–1571.

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

Wang, J., Lu, S., Wang, S.-H., & Zhang, Y.-D. (2022). A review on extreme learning machine. *Multimedia Tools and Applications, 81*(29), 41611–41660. https://doi.org/10.1007/s11042-021-11007-7

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

Zaken, M. van A. (2024, January 17). *Government-wide vision on generative AI of the Netherlands* [Parliamentary document]. Ministerie van Algemene Zaken. https://www.government.nl/documents/parliamentary-documents/2024/01/17/government-wide-vision-on-generative-ai-of-the-netherlands

SEMIOTICS OF SYNTHETIC MEDIA

GIUDITTA BASSANO,[1] ANDREW MCINTYRE,[2]
PIERO POLIDORO[1]

[1] LUMSA University, Department of Human Sciences, Rome, Italy
g.bassano@lumsa.it, p.polidoro@lumsa.it
[2] University of Amsterdam, Department of Media Studies, Amsterdam, the Netherlands
a.mcintyre@uva.nl

This chapter proposes a dual framework for interpreting synthetic media by coupling generative semiotics (enunciation, plastic/figurative isotopies, anchorage, uncanny cues) with Actor-Network Theory (ANT) mapping of production, circulation, and reception. We first situate synthetic images within longer genealogies of manipulation while stressing contemporary discontinuities in scale, speed, access, and political stakes. We then articulate how meaning emerges at two levels: internal textual organization and the socio-technical networks of datasets, models, platforms, policies, and audiences. Four case studies (satire/meme, advertising resurrection, televisual "interview," participatory grotesque) demonstrate how contracts of veridiction shift across genres and contexts. Finally, we introduce a semiotically grounded taxonomy and a reception matrix oriented to political prevention and media-education strategies, privileging capacity-building over mere detection.

DOI
https://doi.org/
10.18690/um.feri.2.2026.3

ISBN
978-961-299-109-8

Keywords:
synthetic media,
semiotics,
actor-network theory,
disinformation,
prevention

University of Maribor Press

1 Digital Media and Actor-Network Theory

Countering deepfakes requires moving beyond technical detection to the narrative coherence and networked conditions that lend synthetic images persuasive force. We therefore combine an external ANT approach - mapping actors from developers to regulators and users - with an internal semiotic approach - tracking isotopies, enunciative positioning, anchorage, and uncanny signals. A genealogical detour clarifies what is continuous with legacy manipulation and what is genuinely new in today's platformed ecologies. The subsequent case synopses function as paradigms, showing how genre, circulation, and audience competence modulate interpretation. On this basis, Section 5 advances a taxonomy of fakeness (expression vs. content) and a four-situation reception model (contract, accident, unmasking, deception) to inform education-first preventive policies.

Before considering the semiotic analysis of deepfakes and synthetic media in their own right, it is important to first consider the broader social environment from which this content emerges and in which it circulates. Such an environment encompasses the development of generative AI systems, considers the distribution of synthetic media across online platforms, and is shaped by policy and legislation. To fully elaborate on the diversity and complexity of this social environment, it is necessary to move beyond traditional socio-technical systems theory (Ropohl 1999), as this cannot fully account for the deep social integration of generative AI systems. Rather, we might better describe this environment through actor-network theory (ANT). Where traditional socio-technical systems theory is somewhat limited to specific systems or contexts in which humans and technology are closely linked (e.g., factories, offices, IT systems), ANT enables SOLARIS to consider a far broader network of social actors involved in the production, dissemination, and reception of synthetic content (e.g., social media users, policy institutions). Furthermore, ANT provides a bridge between socio-technical systems theory and semiotic analysis by highlighting how AI systems contribute to the production of knowledge and how other social actors influence this production. This short section provides only a broad overview of the ANT analysis of the social environment around synthetic media (Bisconti et al., 2024, McIntyre et al., 2025).

Closely associated with the work of theorists such as Bruno Latour, John Law, and Michel Callon, Actor-Network Theory (ANT) is a radical departure from traditional sociology. Rather than focusing on rigid social structures and abstract social forces, ANT conceptualizes any social activity as a dynamic and continually changing network of relationships between different social actors. Importantly, within ANT, a social actor does not solely refer to human beings but further includes a wide range of material entities, including objects, animals, texts, technologies, and institutions. All of these disparate entities are understood to interact with one another within a flat, non-hierarchical network such that every actor, be they human or non-human, can influence the network's dynamics. As these interactions are fluid, the boundaries and exact composition of a social network are never fixed. ANT is not intended as a strict or consistent theoretical framework but, rather, a flexible and evolving approach with its own ambiguities and limitations that even scholars like Latour, Law, and Callon have openly acknowledged (Callon, 1984; Latour, 2007; Law, 1992). That being said, ANT's focus on materiality in social interactions and its inclusion of non-human entities as active social participants means it presents a valuable framework considering the social function of generative AI systems. Of particular interest to SOLARIS' discussion of deepfakes and democracy, ANT allows us to map the vast and diverse network of social actors involved in the production, distribution, and reception of harmful AI-generated content online. This mapping furthermore enables us to understand how socio-political values are introduced and spread throughout this network. As such, we may begin to identify points of policy or legislative intervention to combat democratic risks, which will be discussed in later chapters.

When an internet user views synthetic content online, a particular network of interconnected social actors is formed. This network is expansive and complex, with numerous social actors involved and all linked together by precarious relations. While it is impossible to fully represent such a network, we can develop a simplified version (shown in Figures 1 and 2) in order to identify the key social actors at play, to elaborate on their different characteristics, and to illustrate how these actors are linked within the network. It is important to note that these diagrams are not intended as representations of real-world systems but rather as analytical instruments or provisional maps that allow us to trace associations. These associations have been reconstructed through an ANT-inspired systematic mapping. First, human and non-human actors were identified through a survey and analysis of documentation and sources, including academic literature, policy reports, regulatory texts, journalistic coverage, and publicly available material produced by AI companies. Secondly, the socio-technical chains that link these actors were reconstructed by following the actors themselves and mapping associations from development through circulation to reception.

Where Figure 3.1 provides a basic overview of the different groupings of social actors and how they are typically understood to interact with one another, Figure 3.2 unpacks these groups in more detail. These groupings include social actors involved in the development and distribution of a generative AI system, the creation of synthetic content using these systems, the circulation of this content in online spaces, the user reception of the content, the various policy and legislative interventions, and the broader public discourse surrounding synthetic content. The arrows shown in Figures 3.1 and 3.2 indicate only a possible pipeline of interactions with each social actor impacting upon the next in the sequence. A brief explanation of each stage is provided below.

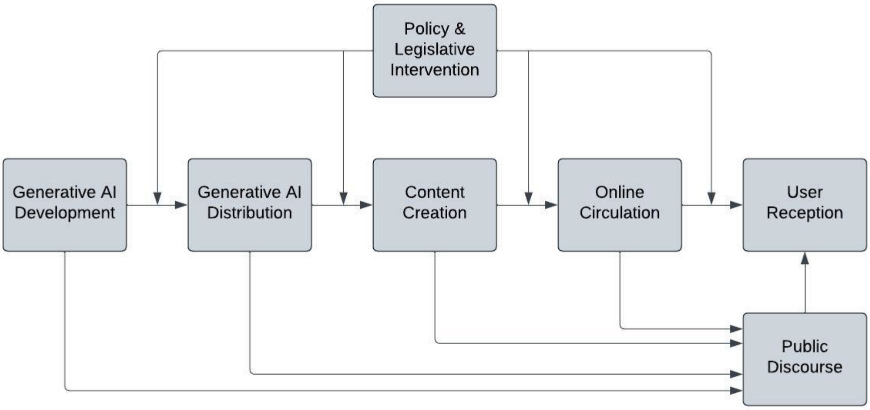

Policy & Legislative Intervention
Generative AI Development
Generative AI Distribution
Content Creation
Online Circulation
User Reception
Public Discourse

Figure 3.1: A general approximation of the significant groups of social actors involved in the production
Source: Bisconti et al., 2024.

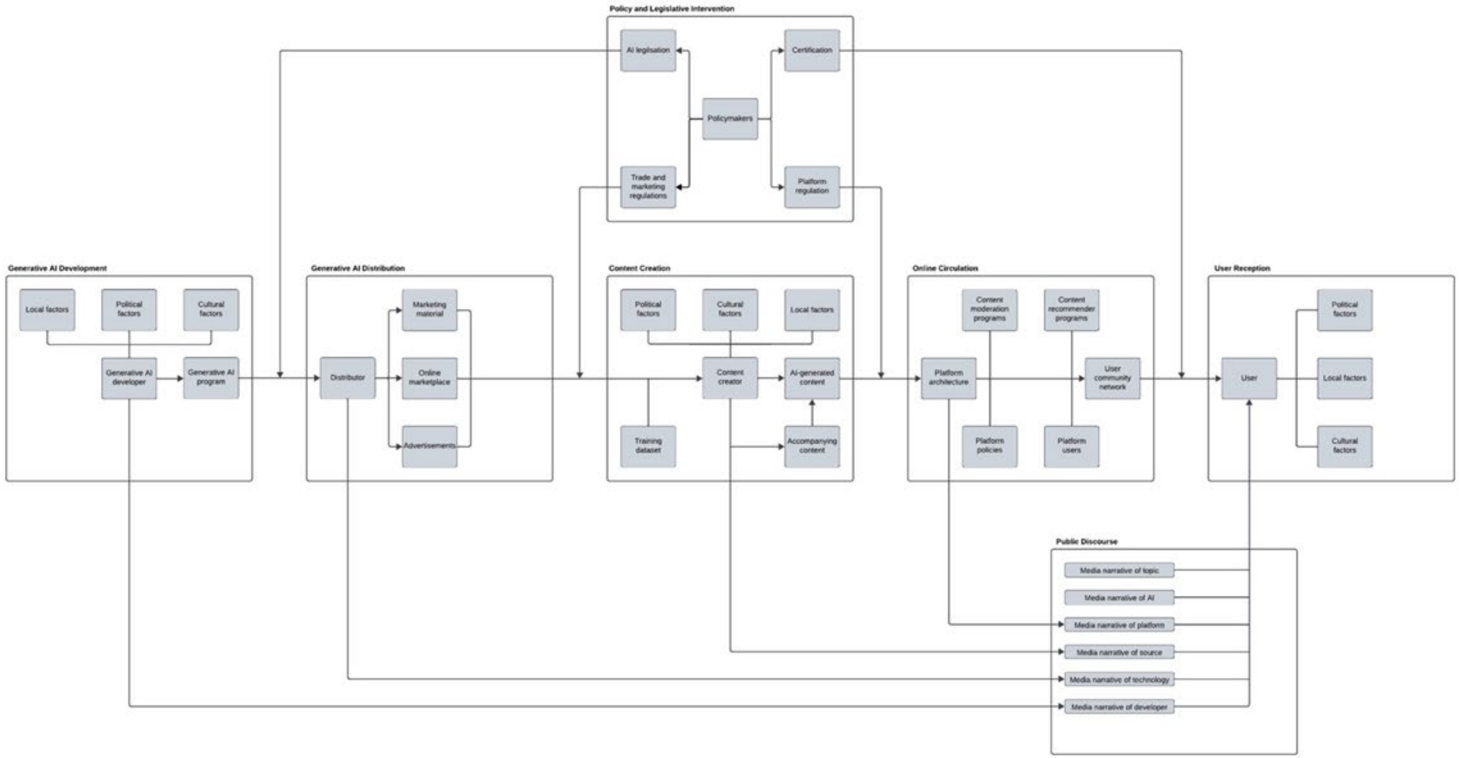


Figure 3.2: An expanded view of the network of social actors involved in the production, circulation and reception of AI-generated content online
Source: Bisconti et al., 2024.

First and foremost, the development of generative AI systems involves diverse actors (e.g., government institutions, private companies, research centres, independent programmers). The design of such technologies is greatly influenced by these actors' motivations (e.g., profit, innovation, public service), access to resources (e.g., researchers, funds, equipment), regulatory compliance (e.g., AI Act, DSA), and adherence to ethical standards (e.g., OECD AI Principles). Furthermore, there are also political, cultural, and local factors that influence these actors and their development processes. Design choices might be shaped through political pressures and public opinion, dominant cultural values, and/or community relations and links to industrial societies. The specific characteristics of these social actors are important to consider as they determine the technical design of an AI system (e.g., datasets, architecture, accuracy, limitations), which, in turn, might lead to bias, inaccuracy, and censorship in the synthetic content generated by these systems. These have socio-political problems. To address this, there are ongoing efforts to introduce value-sensitive design and global initiatives (e.g., UNESCO, NIST) seeking to embed human rights and ethical principles in AI systems at the design stage.

Those social actors involved in the marketing, advertising, and distribution of generative AI systems then further shape how these technologies are perceived and used through promotional materials, advertisements, and visual presentation in online marketplaces. Advertisements and marketing strategies influence who adopts these technologies and for what purposes by encouraging specific uses (e.g., entertainment, pornography) or by appealing to particular user groups (e.g., influencers, programmers). Such practices often embed socio-political values; for example, promoting generative AI technologies for non-consensual pornography perpetuates misogynistic ideas. Meanwhile, hype and exaggeration may misrepresent the technology's capabilities (e.g., reliability, objectivity), thus enabling uncritical or harmful use.

When considering the factors influencing the creation and publishing of synthetic content, it is necessary to account for the content creator's motivations, the kind of synthetic content, and any accompanying material. Whether individuals, groups or institutions, synthetic content creators publish content for certain purposes, including entertainment or disinformation. Their actions are influenced by political, cultural, and local contexts. For example, cultural values (e.g., patriarchal norms) can normalize and encourage content creators to produce exploitative content like deepfake pornography, while unstable or polarized political environments might incentivize political manipulation. The publication of such content itself frames audience interpretation of it as popular, socially acceptable, true or untrue. Deepfakes serve various ends: politically motivated disinformation, ideological reinforcement, or visualization of historical, speculative or political narratives. Such content can undermine institutions, perpetuate biases or reshape public discourse through persuasive synthetic media.

When considering the intended targets of deepfake content, social environments shape their vulnerability and representation. Targets may be individuals, groups, objects, events, or hypothetical scenarios, each carrying characteristics such as demographic profile, societal status, or political significance. Political figures and events are especially at risk. Cultural contexts also influence vulnerability. Celebrities or culturally significant people are attractive targets due to their symbolic value, while misogynistic cultures make women especially susceptible to sexual deepfakes. Deepfakes targeting political figures often misrepresent individuals and their associated organizations and ideologies, amplifying disinformation and undermining broader political movements or institutions.

Social media platforms play a significant role in mediating the dissemination of deepfake content, focusing on their architecture, policies, automated systems, and user interactions. Platform architecture shapes how content is shared and received through features such as newsfeeds, hashtags, trending sections, likes, and comment threads. These design choices frame deepfakes in ways that may obscure their artificiality or amplify their reach. Recommendation algorithms further personalize content delivery, often reinforcing homophily by exposing users to material aligned with their existing interests and values. In the case of deepfakes, this can normalize misleading or polarizing material.

Platform policies and content moderation systems govern which forms of content are allowed, flagged, masked, or removed. Automated moderation programs filter vast amounts of data but are shaped by technical limitations and policy interpretation. These practices intersect with national and international regulations, such as the EU AI Act's transparency requirements for labelling AI-generated content.

Users themselves drive circulation: liking, commenting, and sharing increase visibility, while user networks (e.g., family, friends, colleagues) determine trust and influence. Even users aware of inauthenticity may promote deepfakes for political or ideological reasons.

Finally, social media networks foster "neighbourhoods" or echo chambers, where people cluster by shared identity or opinion. Within these spaces, deepfakes and disinformation can spread quickly with little critique, fuelling polarization and extremism. Efforts to curb harmful content through censorship or labelling may reduce its spread, raise free speech concerns, and push users toward less regulated platforms.

It is not enough to consider synthetic media in isolation. Synthetic content is embedded within wider media ecosystems and shaped by prevailing narratives that influence how it is received and shared. These narratives can relate to the content creator, target, developer, platform, AI technology, or the topic itself. For instance, the perceived trustworthiness, political affiliation, or expertise of a creator can frame how viewers interpret a deepfake. Similarly, targets often carry media personas established through appearances and statements; if a deepfake aligns with or contradicts this persona, it may appear more credible or cause greater reputational damage.

Narratives about AI technology also matter. Some media emphasize deepfakes' inaccuracy, encouraging uncritical acceptance, while others highlight their sophistication, fostering scepticism. This duality impacts the perception of harmful deepfakes and the uptake of pro-democratic applications. Developer and platform identities shape interpretation too: trustworthy brands or platforms with strong moderation may lend legitimacy, while weakly moderated spaces foster doubt.

Broader media coverage of topics featured in deepfakes, such as political controversies, can amplify their impact. Meanwhile, AI "hype" promoted by developers and media often misrepresents capabilities, portraying technologies as neutral and objective. News organizations play a dual role, sometimes debunking disinformation, and especially when under-resourced, unintentionally perpetuating it.

Across all these different social actors and interactions, there may be policy and regulatory interventions in the production, circulation, and reception of deepfake content. Key actors include government officials, regulators, legislation, and certification mechanisms. Policymakers' political affiliations and status shape the form and implementation of policies. AI-specific legislation, such as the EU AI Act, introduces transparency requirements mandating that AI-generated content be labelled, with further laws expected as risks emerge. Trade and marketing regulations govern how AI products are promoted, preventing misleading claims, while platform regulations control how deepfakes circulate, particularly harmful material like pornography.

Certification adds another layer, with fact-checkers labelling false or misleading content, while "pre-bunking" initiatives raise awareness of manipulative techniques, fostering media literacy. Globally, three spheres dominate regulation: the US emphasizes market-driven self-regulation, China enforces state-driven control embedding political values, and the EU adopts a rights-based, transparency-focused model. The EU AI Act exemplifies this, with strict labelling and oversight requirements. Its influence is expected to spread internationally through the "Brussels Effect," setting global standards for ethical AI governance.

Finally, when considering the user themselves and how they receive deepfake content, it is necessary to understand how their personal characteristics and social environments shape their perception. Individual factors include demographics, education, media literacy, knowledge of AI, political affiliation, and societal roles (e.g., journalists, academics, or officials) which can make some users more influential or vulnerable to manipulation.

User environments also play a key role. Political factors, including local policies, pressure groups, and prevailing public sentiments, influence susceptibility, while cultural factors (e.g., ethnic, religious, national, or institutional) shape how content is interpreted. For example, journalists may prioritize sensational content to attract audiences, affecting dissemination.

Many users approach deepfakes uncritically due to the novelty and rapid development of AI, combined with marketing and media hype portraying technologies as objective or authoritative. This can lead users to accept AI-generated content as truthful and adopt the political ideas it conveys, particularly regarding complex or nuanced issues, increasing the risk of manipulation and misinformation.

Ultimately, adopting an ANT perspective enables us to understand GenAI not merely as a set of technologies but as active social actors. This approach reveals the complex social environment in which these technologies operate and how this environment shapes the production and circulation of content, as well as the semiotic meanings embedded within it. By foregrounding these networks of influence, we gain a richer understanding of how GenAI influences social and cultural discourse and values. This ANT mapping functions as an overarching analysis against which a more focused semiotic analysis of specific synthetic images is conducted.

This chapter adopts a dual analytical lens. Internally, each synthetic image is examined through a generative semiotic grid (plastic and figurative isotopies, enunciative configurations, anchorage, uncanny cues). Externally, an Actor-Network Theory mapping identifies the socio-technical actors involved in the image's production, circulation, and reception. The two procedures are applied in parallel, allowing us to link textual micro-coherence to the broader networks of platforms, models, norms, and audiences that shape meaning.

2 Continuities and Discontinuities between Legacy and Synthetic Media

The analysis of *synthetic media* cannot ignore a comparison with previous media traditions. To understand the scope of the transformations underway, it is necessary to distinguish the lines of continuity from the breaks introduced by generative artificial intelligence. Visual manipulation is certainly not a recent invention. As early as the 19th century, photomontage (Floch, 1986) enabled the recombination of image portions to achieve illusionistic or satirical effects. In the 20th century, *airbrushing* and photo editing practices consolidated an imaginary world in which images were never a guarantee of absolute truth. Similarly, political satire has long employed caricature and distortion to challenge the authority of leaders. *Synthetic media* are therefore part of a long genealogy of forms of alteration, spanning photography, cinema, and television. The use of digital CGI techniques in cinema during the 1990s and 2000s can also be considered a precursor: The reconstruction of impossible scenarios and non-existent characters has accustomed viewers to suspend their disbelief and accept simulated worlds as an integral part of collective visual culture.

What has changed radically with *synthetic media* is the speed, scale, and social diffusion of these practices. Whereas in the past manipulation techniques were the preserve of specialists, today accessible tools such as Midjourney, DALL-E, or Veo allow anyone to generate photorealistic images and videos with a simple text prompt. The emergence of the *prosumer*, the user-producer, marks a qualitative leap in the democratization of visual manipulation. Another discontinuity concerns circulation. Legacy media were based on centralized distribution logic (newspapers, television, cinema), while *synthetic media* spread through digital platforms that reward virality, remixing, and participation. Editorial institutions no longer regulate the normativity of public discourse, but by recommendation algorithms and online communities. Another critical aspect in this regard is related to intentionality. Just think, for example, that photo editing in the second half of the 20th century was a practice linked to aesthetic dominance.

In most cases, retouching was equivalent to "perfecting" and "beautifying". Finally, the political stakes are higher. While traditional satire could be easily recognized as such, today a deepfake can be confused with an authentic document and have immediate consequences in terms of reputation, credibility, and even international security. The difficulty of distinguishing between true and false undermines social trust, shifting the focus from objective evidence to subjective beliefs. In summary, *synthetic media* represent a continuation of existing manipulation practices, but they introduce radical discontinuities in terms of accessibility, speed, scale of dissemination, and political impact. Semiotics, in dialogue with the social sciences, must therefore address the tradition of visual falsification and the new ecologies of visibility produced by digital platforms.

3 Semiotic Frameworks for the Analysis of Visual Texts

To understand *synthetic media*, we need to construct a theoretical framework capable of bringing together the internal mechanisms of signification and the socio-technical chains that make its production and circulation possible. In this sense, the convergence between generative semiotics and Actor-Network Theory (ANT) proves particularly fruitful. Methodologically, this chapter combines a visual-semiotic analysis of synthetic images with an Actor-Network Theory mapping of the socio-technical actors that shape their production, circulation, and reception. Semiotics, in the tradition of Greimasian semiotics (Greimas, 1976; Greimas & Courtés, 1979, 1986), offers tools for describing the internal coherence of visual texts. Each image is organized by figurative and plastic isotopies, which establish fields of meaning and orient perception. Figurative isotopies refer to "coherent" and even redundant recurrences of recognisable elements: these recurrences allow us to evaluate, for example, the degree of verisimilitude of a photographic background in relation to what is seen in the foreground.

When we talk about plastic isotopies, we are referring to the consistency between formal elements, such as colours, lines, lighting, and spatial distribution. For example, in an AI-generated photo, we can notice that some of the contours of an object or body part are "blurred" and thus understand that it is an artificial image. In this way, many elements contribute to producing a reality effect: in artificial photos, this can be unmasked more or less easily depending on the observer's interpretative skills.

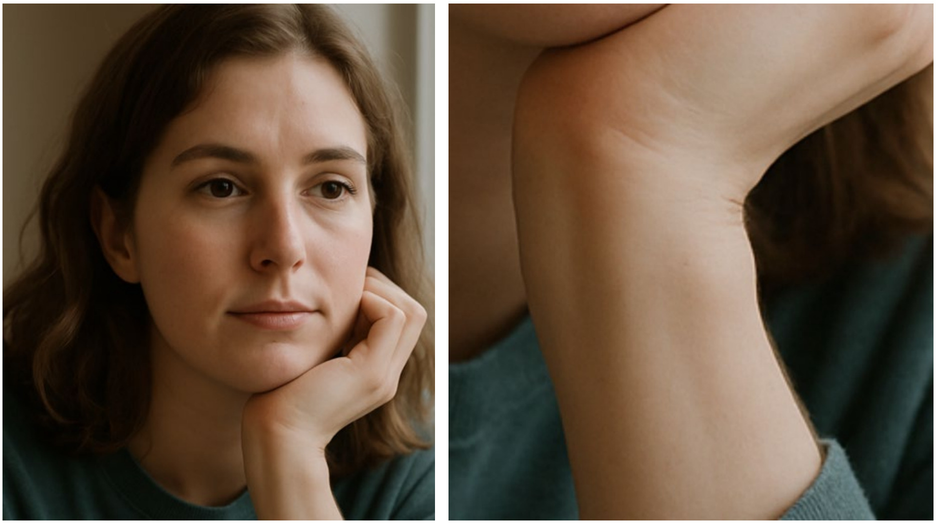


Figure 3.3: In this portrait generated by ChatGPT, a very small detail on the wrist shows an unnatural edge, inconsistent with the normal perception of a human wrist fold.
Source: copyright Giuditta Bassano.

Furthermore, enunciative configurations, as markers of point of view, deictic strategies, and signals of the author's presence/absence contribute to building a communicative contract with the user (Dondero, 2020). In synthetic media, these elements take on even greater significance because their opacity or ambiguity can be easily concealed. For example, there are seemingly credible nature videos circulating in which two nocturnal animals of different species appear to be playing together; however, in reality, they belong to species that do not live in the same climate or on the same continent. The verisimilitude of such videos stems from the fact that they "simulate" the typical aesthetics of infrared LED footage from camera traps used in documentaries.

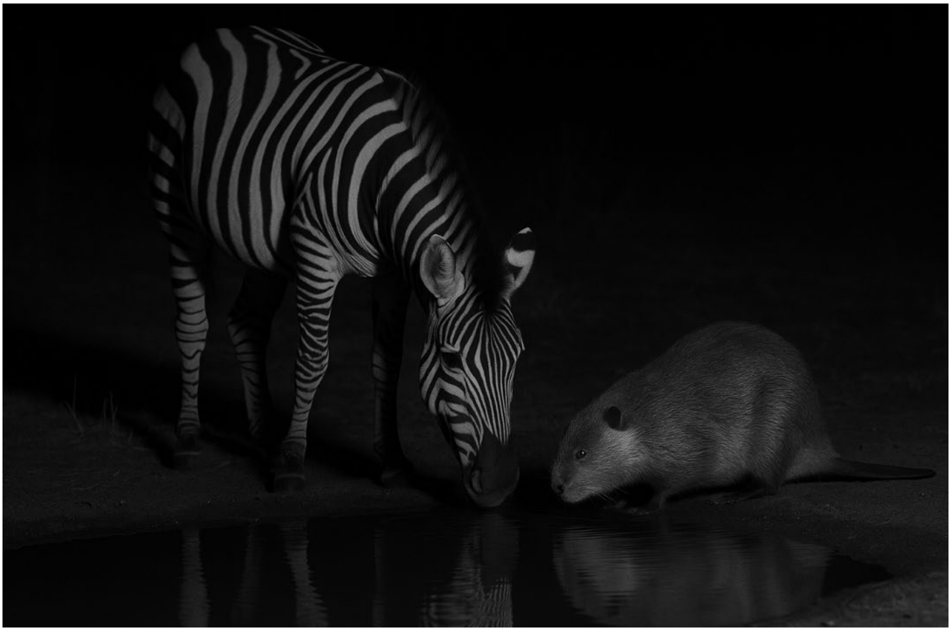


Figure 3.4: The Grant's zebra and the Canadian beaver do not live on the same continent in any way.
Source: Photo generated by ChatGPT, prompt by Giuditta Bassano.

Finally, there is also a phenomenological-semiotic problem: namely, the way in which our perception seeks to "find" a principle of humanity in objects, in moving shapes, and in toys - consider the phenomenon of *pareidolia* (Eco, 2010). A case in point is the so-called *uncanny valley* (Leone, 2021): when a synthetic face is almost realistic, but not quite, the observer recognizes the artificial nature of the face, but at the same time continues to receive an intermittent impression of humanity. Thus, a semiotics of the uncanny (Kress & Leeuwen, 2020; Leone & Gramigna, 2021) allows us to analyse these micro-clues of non-humanity as inconsistent isotopies that undermine the effect of verisimilitude. The phenomenon is not limited to physiognomy but can emerge in environmental details and bodily postures.

Semiotic analysis, therefore, does not seek to technically unmask the algorithm, but rather to reconstruct how signs of artificiality translate into meaning for different audiences. At the same time, ANT allows us to place these texts within broader socio-technical networks. Indeed, a deepfake never exists in isolation: it is the product of complex chains that include generative model developers, training datasets, distribution platforms, content creators, moderation policies, fact-checkers, and end users. The analysis of a synthetic visual text must, therefore, be articulated on two complementary levels: on the one hand, the internal semiotic organization, and on the other, the translations and mediations carried out by non-human agents (software, algorithms, interfaces) and human agents (authors, institutions, user communities). Verbal anchoring, already described by Roland Barthes in relation to photography, assumes a crucial role here (Leone, 2021). In social media, synthetic images are almost always accompanied by texts: descriptions, hashtags, comments, and captions. These elements not only guide interpretation but can also conceal or reveal the artificial nature of the content. A deepfake declared as parody activates ironic isotopies and is interpreted in a satirical key; the duplicate content, without a label, can be perceived as proof of an event that never happened. The question of anchoring is thus intertwined with the algorithmic logic of visibility and the media normativity of the platform.

4 Critical Case Studies: Deepfakes and Their Semiotic Implications

To gain a deep understanding of the cultural and political implications of *synthetic media*, it is not enough to analyse the phenomenon in the abstract: it is necessary to study concrete cases that serve as litmus tests for the transformations taking place.

The four case studies were selected through a paradigmatic sampling logic rather than by representativeness. Each case illuminates a distinct semiotic and socio-technical configuration: (i) Pope Balenciaga exemplifies hybrid satire and ambiguous veridiction; (ii) Lola Flores foregrounds posthumous identity reconstruction and commercial appropriation; (iii) Dalida highlights televisual enunciation and the redefinition of documentary authority; (iv) the Will Smith meme series captures the rapid evolution of synthetic aesthetics from grotesque error to infrastructural realism. These cases were chosen because they activate different combinations of textual isotopies, uncanny cues, platform dynamics, and actor-network relations, allowing for a comparative framework capable of tracing broader cultural transformations.

4.1 Pope Balenciaga (2023, Midjourney)

The case of the so-called *Pope Balenciaga*, a series of images of the pontiff dressed in a designer white down jacket, generated with Midjourney and circulated online in March 2023, exemplifies the functioning of deepfakes as hybrids between satire and photorealism.

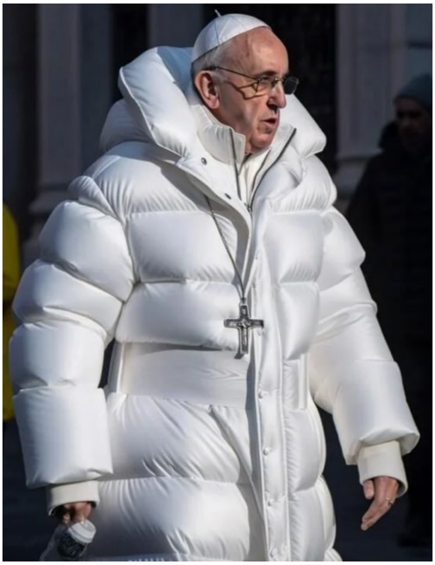


Figure 3.5: One of the most famous artificial images of contemporary times involving Pope Francis.
Source: widely circulated AI-generated image depicting Pope Francis in a white puffer coat.

Internal semiotic analysis:

- Figurative isotopies: the papal white blends with the bright white of the catwalk down jacket; the outfit evokes both ecclesiastical austerity and fashion glamour.
- Plastic isotopies: contrast between the neutral background and the brightness of the garment, which amplifies the effect of hyper-reality.

- Enunciation: the absence of markers of irony within the image generates ambiguity. It is the viral context (memes, ironic comments) that disambiguates.

ANT and socio-technical chain:

- Non-human agents: Midjourney as a platform, a dataset of religious and fashion images.
- Human agents include Reddit and Twitter users who share, journalists who repost, and fact-checkers who clarify the falsehood.
- Effect: oscillation between irony and misinformation, with risks to credibility among visually illiterate audiences.

The case demonstrates how a synthetic image can integrate into a traditional discursive regime (political satire), with its effects amplified by its verisimilitude.

4.2 Lola Flores for Cruzcampo (2021, hybrid media)

Cruzcampo's advertising campaign, which digitally resurrects Andalusian singer Lola Flores in 2021, is an example of *media hybridization*: deepfakes, sound archives and advertising editing converge in a commercial product.

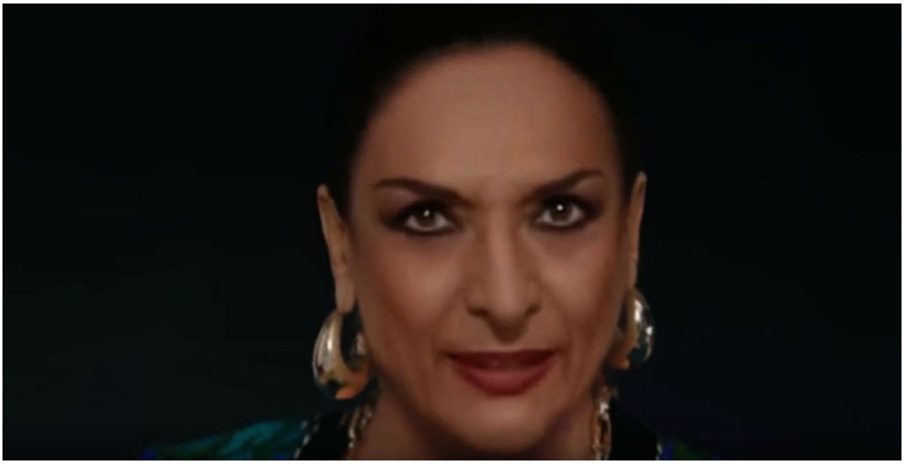


Figure 3.6: A frame from the commercial that digitally resurrects the Andalusian star Lola Flores.
Source: screenshot from "Anuncio Cruzcampo Lola Flores 2021 (Spot TV 30s)", YouTube, JaviTV, January 24 2021.

Internal semiotic analysis:

- Identity isotopies: Flores' reconstructed face becomes a guarantee of authenticity for a message linked to "identidad andaluza" (Andalusian identity).
- Enunciation: the use of the first person ("¿Y tú, sabes quién eres?") creates an effect of proximity that reinforces the emotional impact.
- Uncanny: the body appears alive, but the awareness of the artist's death produces cognitive friction.

ANT and socio-technical chain:

- Non-human agents: face reenactment software, audiovisual archives.
- Human agents include advertising agencies, family heirs (who have given their consent), as well as television and social media audiences.
- Normative dimension: the issue of posthumous consent and the 'delegated responsibility' of the heirs.

This case shows how *synthetic media* can be exploited by the market, transforming cultural memory into an economic resource, with the risk of reducing collective identities to visual commodities.

3.1.1. Dalida in Hotel du Temps (2022, hybrid media)

The French television programme *Hotel du Temps*, hosted by Thierry Ardisson, resurrected deceased celebrities (including Dalida) to 'interview' them in the studio using face-swapping and voice-cloning techniques.

Internal semiotic analysis:

- Enunciation: the 'truth contract' typical of television journalism is grafted onto a digital artifice. The television mise en scène simulates a live interview, blurring the genres of documentary, fiction and talk show.
- Uncanny effect: the viewer oscillates between nostalgic fascination and ethical unease.

ANT and socio-technical chain:

- Non-human actants: face swap software and archived video dataset.
- Human actors: Ardisson as author, digital technicians, and traditional television audience.
- Political effect: redefinition of collective memory, risk of 'affective revisionism' (resurrections that rewrite history).

The Dalida case raises profound questions about posthumousness and the use of images as 'heritable assets' in the absence of clear legislation (Bassano & Cerutti, 2024).

4.4 Will Smith Meme (2025, Veo 3)

The "Will Smith eating spaghetti" meme (2023) and the Veo 3 "Will Smith" frame (2025) (Fig. 6. below) encapsulate the accelerated evolution of generative media aesthetics. The grotesque distortion of the first phase and the photorealistic perfection of the second can be seen as sequential stages of the same cultural experiment: the former tests the limits of plausibility through excess, while the latter redefines plausibility itself as the ultimate aesthetic value.

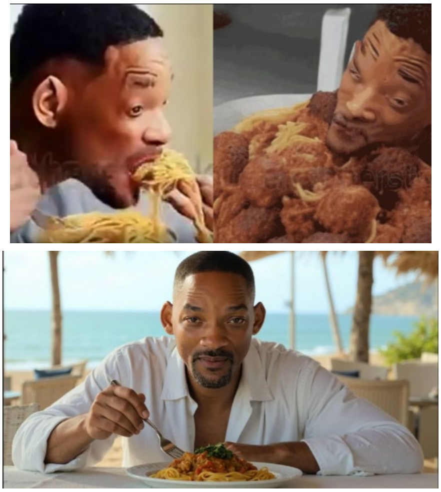


Figure 3.8: The 2023 meta-digital meme Will Smith eating spaghetti becomes, two years later, a temporal anchor for observing an extraordinarily rapid technical evolution.
Source: from the top, screenshot from viral AI-generated deepfake depicting Will Smith eating spaghetti (YouTube 2022, www.youtube.com/watch?v=vbWe5k4fFWE), and screenshot from viral AI-generated deepfake depicting Will Smith eating spaghetti (YouTube 2025, www.youtube.com/@agx_agi).

Figurativeness and isotopies:

In 2023, forms collapse and textures blend: the edible and the human merge into a chaotic visual loop where humour depends on error. In 2025, all plastic elements align - light, texture, colour produce a seamless *reality effect*. The grotesque gives way to the algorithmic normality of lifestyle realism, where perfection itself becomes suspect.

Enunciation and the uncanny:

The 2023 meme was openly parodic, its enunciation collective and self-aware; the Veo 3 image instead speaks as if real, erasing irony and testing the viewer's interpretive vigilance. The uncanny shifts from failure to success: not the deformity of form, but its flawless credibility now unsettles perception.

ANT and socio-technical chain:

From early chaotic engines to Veo 3's multimodal coherence, the generative system evolves from collective play to infrastructural realism. Users move from active co-authors to passive spectators, while platforms reward aesthetic smoothness over disruption. The result is a new threshold of synthetic verisimilitude, where realism itself becomes the message.

Interpretive significance:

Between 2023 and 2025, generative imagery moves from the grotesque to the post-ironic, from visible artifice to imperceptible simulation. What was once laughable for its failure now compels attention for its precision. This shift defines a new mode of spectatorship, grounded not in visual trust but in interpretive literacy – the ability to discern the social and technical networks behind the image.

A comparative reading of the four cases highlights a progressive transformation in the semiotics of synthetic media:

- From irony to transparency: while early cases such as *Pope Balenciaga* relied on ambiguous irony to generate meaning, the *Veo 3 Will Smith* image shows how hyperrealism now erases the ironic frame, demanding new interpretive vigilance.

- Evolution of the veridiction contract: deepfakes increasingly occupy the grey area between fiction and documentation. The advertising and televisual examples (*Lola Flores, Dalida*) demonstrate how synthetic media inherit the authority of their original genres while subtly redefining their truth regimes.
- Ethical and normative complexity: questions of consent, posthumous agency, and delegated authorship move to the foreground, exposing the inadequacy of existing legal and ethical frameworks to manage hybrid human–machine authorship.
- Reconfiguration of participation: from the collective remixing of the 2023 meme to the infrastructural realism of 2025, the human role shifts from playful co-creation to critical spectatorship within algorithmic ecosystems.

Together, these cases map the passage from visible artifice to imperceptible simulation, revealing how deepfakes evolve from cultural anomalies to structural components of media experience. The integration of semiotic analysis and Actor-Network Theory proves essential to understanding this shift, linking textual micro-coherence to the wider networks of production, circulation, and regulation.

5 A Semiotic Framework for Political Prevention

The preceding sections have outlined a theoretical trajectory that moves from the analysis of socio-technical networks (ANT) and media genealogies to the development of a semiotic framework capable of interpreting deepfakes as complex cultural texts. Through this dual perspective – external and internal – it has been shown that synthetic media function not merely as technological devices but as genuine social actors that reshape truth contracts and digital citizenship practices. Section 5 builds on this continuity by proposing an applied interpretive model of deepfakes, translating the preceding theoretical insights into a tool for designing educational policies and interpretive literacy strategies aimed at prevention and democratic resilience. A regulatory intervention should begin with the clearest possible understanding of the subject to be regulated. A helpful way to fix that knowledge is a *taxonomy*: a (more or less) hierarchical set of labels and definitions that lets us relate individual phenomena to broader categories. The act of assigning a single case to a category is called *classification*, and it is always a compromise. On the one hand, we have to downplay (that is, set aside) the case's unique features, which are inevitably lost; on the other, we gain the benefit of placing a single, relatively new phenomenon within a familiar framework that indicates some of its properties (shared with other phenomena) and, ideally, offers practical guidance on how to respond to it.

In this section, then, we aim to give a taxonomic backdrop to the discussion of deepfakes. To do that, we first need to define a few concepts directly or indirectly linked to deepfakes, starting with *fake news* and *post-truth* (Polidoro, 2008).

In short, *post-truth* refers to a supposed shift in contemporary public debate in which emotional factors increasingly outweigh rational ones, and truth matters less than other considerations such as personal or partisan interest. Framed this way, post-truth is a general attitude to truth, and a cultural change located in our present, following the digital revolution. In a post-truth environment, the spread of false reports becomes structural rather than exceptional. Two key terms are *disinformation* (the deliberate spread of false information) and *misinformation* (the unintentional spread of false information that the sender believes to be true). The difference between them lies wholly in the sender's intention to circulate something they know is false. A related expression is *malinformation*: the spread of accurate information with the aim of harming someone (as in gossip). Since malinformation deals with true information, we do not consider it here.

Within this context, *fake news* is central, and part of the deepfake phenomenon can be placed under it. It should make it clear that the label *deepfakes* is misleading, though, because "fakes" suggests something falsified and intentionally produced to deceive. Whereas, as this book has noted repeatedly, not all deepfakes serve this purpose: their synthetic nature can be made explicit, and they can also be used for constructive and positive ends.

The term *fake news* also poses a practical problem: it is an umbrella term that covers many different phenomena. We therefore need to give it an internal structure, namely, a taxonomy of fake news.

The literature offers several attempts at such a taxonomy (Chong and Choy 2020; Jaster and Lanius 2018; Rastogi and Bansal 2022; Tandoc, Lim and Ling 2018; Wardle 2016, 2017), though not many, because attention soon turned to taxonomies of *classification systems*: the (almost always automated) tools used to identify fake news. For a fuller discussion of taxonomies of fake news, see Polidoro 2025. Here, it is worth noting that existing models suffer from two main limits. First, a few models rely – albeit in different ways – on two dimensions: *facticity* (how close, or rather how far, items are from the truth) and the sender's *intention* (for example, parody and satire openly distort reality or construct a non-truthful one). The difficulty lies above all in the latter: intention is interesting, but hard to verify. Second, other models lack system: rather than deriving types from clearly defined dimensions (for example, by combining them), they amount to unstructured lists of different phenomena.

To overcome these limits, the SOLARIS project developed two models grounded in semiotics. They have different aims and viewpoints. For further details, see Polidoro 2025.

The aim of the first model is to build a taxonomy of fake news, which also helps identify different kinds of malicious deepfakes. It does not propose new types; instead, it reorganizes them according to a semiotically grounded logic that differs from what is often found in the literature.

According to this model, we first distinguish fake news produced by falsifying the *level of expression* from those produced at the *level of content*. In the former (which includes deepfakes), falsification acts on the *material form* – visual, audio, or otherwise. This may work on pre-existing material (*manipulation*) or start from scratch (*fabrication*). By contrast, working at the content level means we are not falsifying the vehicle of information (the expression), but, in some way, the content it carries. This can happen in two ways. The first is to create an entirely untruthful report from scratch: *invention*. The second is to manipulate content that is partly or wholly true so that it leads to a mistaken reading of reality. Such manipulation may occur within the text (for instance, through misleading adjectives), between the text and its accompanying elements – the *paratext* (for example, giving a truthful report a skewed headline or pairing it with an image that steers the reader to a wrong interpretation), or between the item as a whole (text, title, image, etc.) and the context in which it appears (for example, placing it alongside other items so that it is framed in a particular way). Finally, the falsehood of a news item can depend not just on the falsity of the content, but also on falsifying the *source* (as when a fake television newscast is produced).

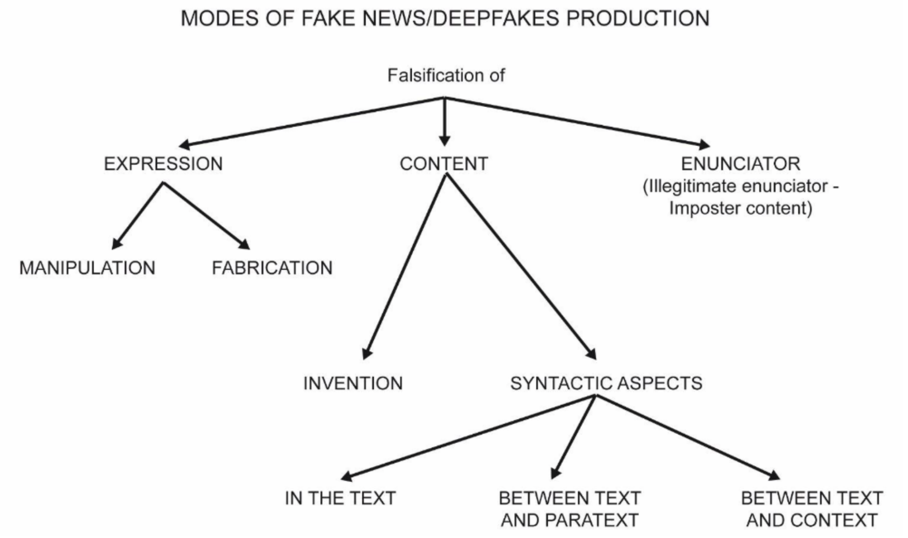


Figure 3.9: Graph depicting taxonomic component characterising different modes of fake news/deepfakes production.
Source: copyright Piero Polidoro.

The diagram above shows how these differences fit together. Because many of these aspects can co-exist within one piece of fake news, the types should not be treated as mutually exclusive. The best way to apply the model is therefore a coding sheet on which to note which taxonomic components appear in each individual item.

The second model sets out the different situations one may face when dealing with fake news. It combines two dimensions. The first concerns the sender, but not their intention (which is hard to prove). Rather, it asks whether the text includes markers that make its fabrication explicit – for example, paradoxical cues (as in parody) or technical ones (such as watermarks). The second dimension concerns the recipient's ability to judge the text's truthfulness. This yields four situations:

- *Contract*: the recipient correctly recognizes a text that is explicitly false (for example, realizing they are engaging with parody).
- *Accident*: through inattention or limited literacy, the recipient fails to recognise an explicitly false text and takes it to be true (as happened with Orson Welles's 1938 radio broadcast of War of the Worlds).
- *Unmasking*: the recipient detects the attempt to deceive and unmasks the fake news.

– ***Deception***: the fake news succeeds in misleading the recipient.

Contract is not problematic, and *Unmasking* is a case of successful, autonomous debunking. The problematic cases are *Accident* and *Deception*. To limit these, we must adapt different strategies: safeguard measures to avoid the former, and capacity-building to strengthen debunking in the latter.

Table 1: Model showcasing four different fake news scenarios faced by senders and users.

		Sender	
		Explicit fabrication	**Implicit fabrication**
Receiver	**Recognition**	avoid ↓ Contract	foster ↑ Unmasking
	Lack of recognition	Accident	Deception

Source: copyright Piero Polidoro.

6 Concluding Remarks

Synthetic media are not a mere by-product of technology but a laboratory of veridiction, where boundaries between truth/falsehood and human/artificial are continually renegotiated. A combined semiotics and ANT lens shows that meaning arises from the interplay of textual micro-cues (isotopies, enunciation, anchorage, uncanny) and macro-structures (models, platforms, norms, audiences). The cases confirm that effects depend less on tools than on discursive contracts, networks of actors, and audience competence. Accordingly, prevention should prioritise interpretive capacity-building: semiotic literacy, transparent labelling regimes, and context-aware pedagogy that reduces *accidents* and strengthens *unmasking*, rather than relying solely on detection. The practical instruments proposed in Section 5 - a taxonomy of fakery and a reception matrix seek to offer a shared vocabulary for scholars, educators, and policymakers to design education-led, democracy-supporting responses to synthetic images.

End notes

Giuditta Bassano and Andrew McIntyre conceptualized the chapter and coordinated the writing. Giuditta Bassano wrote the Introduction, the Conclusion, and the following Sections: "Continuities and Discontinuities between Legacy and Synthetic Media", "Semiotic Frameworks for the Analysis of Visual Texts", and "Critical Case Studies: Deepfakes and Their Semiotic Implications". Andrew McIntyre wrote "Digital Media and Actor-Network Theory", while Piero Polidoro authored the section on "A Semiotic Framework for Political Prevention". All authors reviewed and approved the final version.

References

Barthes, R. (1964). Rhétorique de l"image. *Communications, 4*(1), 40–51. https://doi.org/10.3406/comm.1964.1027

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. https://doi.org/10.1007/s11196-023-10067-2

Bisconti, P., McIntyre, A., & Russo, F. (2024). Synthetic socio-technical systems: Poiêsis as meaning making. *Philosophy & Technology, 37*(3), Article 94. https://doi.org/10.1007/s13347-024-00778-0

Callon, M. (1984). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. *The Sociological Review, 32*(1, Suppl.), 196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Chong, M., & Choy, M. (2020). An empirically supported taxonomy of misinformation. In K. Dahir & R. Katz (Eds.), *Navigating fake news, alternative facts, and misinformation in a post-truth world* (pp. 117–138). IGI Global. https://doi.org/10.4018/978-1-7998-2543-2.ch005

Dondero, M. G. (2020). *The language of images: The forms and the forces.* Springer International Publishing. https://doi.org/10.1007/978-3-030-52620-7

Eco, U. (Ed.). (2010). *The limits of interpretation* (1st Midland Book ed.). Indiana University Press.

Floch, J.-M. (1986). *Les formes de l"empreinte: Brandt, Cartier-Bresson, Doisneau, Stieglitz, Strand.* P. Fanlac.

Greimas, A. J. (1976). *Sémiotique et sciences sociales.* Seuil.

Greimas, A. J., & Courtés, J. (1979). *Sémiotique: Dictionnaire raisonné de la théorie du langage.* Hachette.

Greimas, A. J., & Courtés, J. (1986). *Sémiotique. 2: Compléments, débats, propositions.* Hachette.

Jaster, R., & Lanius, D. (2018). What is fake news? *Verus, 127*(2), 207–223. https://doi.org/10.14649/91352

Kress, G. R., & van Leeuwen, T. (2020). *Reading images: The grammar of visual design* (3rd ed.). Routledge. https://doi.org/10.4324/9781003099857

Latour, B. (2007). *Reassembling the social: An introduction to actor-network-theory* (Paperback ed.). Oxford University Press.

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems Practice, 5*(4), 379–393. https://doi.org/10.1007/BF01059830

Leone, M. (2021). *Volti artificiali / Artificial faces. Lexia, 27*, 1–645.

Leone, M., & Gramigna, R. (2021). Special issue: Cultures of the face. *Sign Systems Studies, 49*(3–4).

McIntyre, A., Conover, L., & Russo, F. (2025). A network approach to public trust in generative AI. *Philosophy & Technology, 38*(4), Article 137. https://doi.org/10.1007/s13347-025-00974-6

Polidoro, P. (2008). *Che cos'è la semiotica visiva* (1st ed.). Carocci.

Polidoro, P. (2025). Two proposals for a semiotic taxonomy of fake news and deepfakes. Actes Sémiotiques, (133). https://doi.org/10.25965/as.8964

Rastogi, S., & Bansal, D. (2023). A review on fake news detection 3T's: Typology, time of detection, taxonomies. *International Journal of Information Security, 22*, 177–212. https://doi.org/10.1007/s10207-022-00625-3

Ropohl, G. (1999). Philosophy of socio-technical systems. *Techne: Research in Philosophy and Technology, 4*(3), 186–194. https://doi.org/10.5840/techne19994311

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news." *Digital Journalism, 6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Wardle, C. (2016, November 18). 6 types of misinformation circulated this election season. *Columbia Journalism Review*.https://www.cjr.org/tow_center/6_types_election_fake_news.php

Wardle, C. (2017). Fake news: It's complicated. *First Draft*. https://firstdraftnews.org/fake-news-complicated/

The Psychology of Deception: Why We Believe Deepfakes

Nejc Plohl,[1] Urška Smrke,[2]
Letizia Aquilino,[3, 4] Izidor Mlakar[2]

[1] University of Maribor, Faculty of Arts, Maribor, Slovenia
nejc.plohl1@um.si
[2] University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
urska.smrke@um.si, izidor.mlakar@um.si
[3] DEXAI – Artificial Ethics, Rome, Italy
[4] Università Cattolica Del Sacro Cuore, Milan, Italy
letizia.aquilino@dexai.eu

DOI
https://doi.org/10.18690/um.feri.2.2026.4

ISBN
978-961-299-109-8

Advances in artificial intelligence have enabled the creation of highly realistic deepfakes, yet their impact ultimately depends on how humans perceive and interpret them. This chapter examines the psychological processes underlying belief in deepfakes, focusing on perceptual mechanisms, individual differences, and downstream consequences. Despite widespread confidence in detection abilities, people generally struggle to distinguish authentic from manipulated videos, often performing at or near chance. To move beyond binary detection measures, we introduce the construct of perceived trustworthiness, defined as the extent to which a video is experienced as authentic. We describe the development and validation of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ), which captures two dimensions: trustworthiness of content (plausibility and source credibility) and trustworthiness of presentation (perceived realism of delivery, including technical quality, voice, and behaviour). This tool enables systematic examination of perceptual features that make deepfakes believable across contexts. We further show how sociodemographic, motivational, and cognitive factors shape susceptibility, and demonstrate that perceived trustworthiness predicts attitudes toward climate change and immigration as well as intentions to share content. Overall, the chapter highlights the need for psychological, not only technological, interventions.

Keywords:
deepfakes,
psychology,
trustworthiness,
individual differences,
attitudes

University of Maribor Press

1 Introduction

While chapter 1 introduced the technical layers of deepfakes, explaining how advances in artificial intelligence, machine learning, deep learning, and generative adversarial networks make it possible to create hyper-realistic synthetic media, technology is only half of the story. The other half lies in human perception, specifically in how we see, interpret, and, in the end, decide whether to believe what is placed before our eyes and ears. No matter how sophisticated a deepfake's creation process is, its final impact depends on the processes occurring within the person encountering it. However, these perceptual and cognitive processes depend on broader individual characteristics and are particularly complex in the context of multimodal media, making it difficult to fully grasp why people come to believe deepfakes.

Specifically, how we judge a video's authenticity is not shaped solely by the sensory information it provides, but also by who we are as individuals, our prior knowledge, worldviews, cognitive styles, and even habitual media use (Somoray et al., 2025). Two people can watch the same deepfake and come away with very different conclusions, depending on factors such as political orientation, trust in institutions, or media literacy. This highlights the importance of individual differences, which interact with perceptual processes to shape how a given deepfake is received and interpreted.

Second, the challenge is compounded by the fact that deepfakes are multimodal, targeting several channels of human perception simultaneously (Lee & Shin, 2022). They can look real, sound real, and convey a message we are already predisposed to accept. This convergence of visual, auditory, and semantic cues can create a powerful sense of authenticity, making it harder for viewers to engage in critical evaluation. Even when technical imperfections are present (e.g., slightly unnatural facial movements, subtle audio mismatches), a coherent and plausible message can override scepticism, fostering misplaced but compelling trust. Understanding how these different pathways interact, and how they are related to individual characteristics, is essential for building a comprehensive account of why people believe deepfakes and how they can be influenced by them. Crucially, such influence is not limited to the moment of exposure; perceptions of authenticity can shape downstream psychological outcomes (Rijo & Waldzus, 2023), including changes in attitudes toward the depicted topic and intentions to engage with or share the content. These behavioural consequences, ranging from private opinion shifts to the viral spread of misinformation, make the study of deepfake perception a matter of detection accuracy and of understanding their broader persuasive power.

In the present chapter, we hence focus on the human aspect of deepfakes, with a particular emphasis on the advancements made within the SOLARIS project (which are presented in detail in our research articles; Plohl et al., 2024, 2025a, 2025b, 2025c). We start by reviewing the key literature on human detection of deepfakes. Next, we move beyond detection and introduce the concept of perceived trustworthiness of deepfakes to provide some insight into the perceptual elements of deepfakes that make people more or less inclined to believe them. We then accompany these perceptual aspects with broader individual characteristics, which may contribute to individuals' susceptibility to deepfakes. Lastly, we finish the chapter with a brief section on why deepfake detection and perceived trustworthiness matter. Altogether, the chapter provides a brief but comprehensive insight into the psychological processes underlying how people perceive and respond to deepfakes, highlighting both perceptual and individual factors that shape susceptibility and resistance.

2 Do We Actually Believe Deepfakes?

People generally believe that they can reliably detect deepfakes and overestimate their performance in deepfake detection tasks (e.g., Köbis et al., 2021; Somoray & Miller, 2023), which is particularly true for those who actually perform the worst in such tasks (Plohl et al., 2025c), illustrating a phenomenon called the Dunning-Kruger effect (Kruger & Dunning, 1999). However, in reality, the existing studies suggest that we are generally bad at recognizing whether the video is real or manipulated. For example, Köbis and colleagues (2021) exposed participants to 16 videos lasting about 10 seconds and found the overall accuracy level to be 57.6%, just slightly above what would be achieved with coin-tossing (50.0%). Similarly, another recent study (Somoray & Miller, 2023) found the mean categorization accuracy of 20 videos lasting 10 seconds to be 60.7%, which, again, only slightly exceeded chance levels. Moreover, our recently conducted study revealed that detection accuracy varies based on deepfake quality, manipulated by (mis)aligning the content of the message with the depicted person's actual stance on the topic and changing the technical proficiency (e.g., voice quality, lip-syncing). In this study, 43.5-60.4% of individuals correctly identified lower-quality deepfakes (characterized by misaligned content and low technical proficiency), whereas higher-quality deepfakes (characterized by aligned content and high technical proficiency) were correctly detected only by about a third of participants (30.9-36.6%; Plohl et al., 2025b).

The findings of individual studies have recently been summarized in a comprehensive systematic review investigating deepfake detection. Diel and colleagues (2024) synthesized the evidence on the human ability to detect deepfakes of different modalities, including audio, image, and video. They found 56 studies involving more than 86,000 participants that involved some kind of deepfake stimuli and detection performance measures (which varied between the studies). They found the total deepfake detection accuracy of 55.5% (audio: 62.1%, images: 53.2%, video: 57.3%), which is not significantly above the chance level. Similar results emerged for other metrics beyond analyses of proportions. Hence, the available evidence suggests that individuals' decisions regarding video authenticity are close to decisions one would make by blind guessing, with detection accuracy likely facing additional challenges once deepfakes become more and more sophisticated.

3 Moving Beyond Detection to Understand Why We Believe Deepfakes

Focusing solely on detection and employing simple dichotomous questions asking whether a video is real or a deepfake offers an interesting insight into the extent to which people may believe deepfakes. However, such research cannot convincingly answer how these judgments are formed, or, in other words, why people believe deepfakes. To address this gap, we proposed a new construct, "perceived trustworthiness of deepfakes', defined as the extent to which individuals perceive deepfakes as authentic (i.e., not fabricated). From the beginning, perceived trustworthiness was hypothesized to be multidimensional, consisting of various aspects that may contribute to deepfakes being perceived as more or less trustworthy. Due to specific aspects determining these perceptions not being well-understood and the lack of measures capable of capturing this newly-proposed construct, we set out to develop a new scale by employing a complex process combining various methodologies (i.e., qualitative and quantitative research), stakeholders (i.e., experts and general population), and cultural backgrounds (i.e., participants from the United Kingdom, Italy, and Slovenia).

Specifically, the development and validation of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ; Plohl et al., 2024) occurred in three phases to ensure the scale's validity and conceptual depth. The first phase was dedicated to the development of the initial pool of items. We reviewed the literature to collect items from existing relevant scales (e.g., Hameleers et al., 2024; Hwang et al., 2021; Lee & Shin, 2022) and generate new items based on aspects identified as important in previous, mostly qualitative, studies, such as blurriness on the eye region, abnormal mouth movements, and unnatural voice (e.g., Hameleers et al., 2023; Tahir et al., 2021; Thaw et al., 2021). Furthermore, we conducted face-to-face interviews with students and an online survey with citizens, journalists, and experts. In both interviews and the online survey (overall *N* = 26), participants were asked to watch multiple videos, some of which were deepfakes, decide whether they trust each of them, and share all the thoughts that popped into their heads while forming these decisions. The relevant statements collected qualitatively were transformed into questionnaire items. Lastly, we generated additional items using the Psychometric Item Generator (Götz et al., 2023), a machine-learning solution to developing items for psychometric scales. Altogether, the first phase resulted in 419 initial items.

After reducing the number of items by only keeping those that were unique and general enough (i.e., suitable for different deepfake videos), 123 items were reviewed by 13 experts for content validity. Specifically, the experts were asked to assess the relevance and clarity through a classic content validity procedure. For each item, we then calculated the content validity ratio (a measure of relevance) and content validity index (a measure of clarity), with only items above the acceptable thresholds being retained further. This procedure resulted in a 31-item version covering key dimensions such as the content of the video, the behaviour of the person in the video, the video's source, and its technical features. The items were then translated into Italian and Slovene using the translation-back translation procedure.

In the last step, we conducted large-scale surveys across English, Italian, and Slovene samples (*N* = 733) to investigate the factorial structure of the questionnaire, measurement equivalence of the three language versions, internal reliability of the questionnaire, construct validity, and incremental validity. The results of exploratory and confirmatory factor analyses supported a two-factor structure of the final 22-item scale, consisting of perceived trustworthiness of content (i.e., evaluations of the presented information and its source; 11 items) and perceived trustworthiness of presentation (i.e., evaluations of how the information is presented, including the speaker's behaviour and the video's technical sophistication; 11 items). For instance, a deepfake of a politician delivering factual information aligned with what they usually advocate for may score high on content trustworthiness but low on presentation trustworthiness if the lip-syncing is misaligned. In addition, we found support for configural and metric invariance across the three languages, suggesting that the factor structure and factor loadings are similar across different versions of the questionnaire.

The scale demonstrated strong psychometric properties, including high reliability (α = .83–.92). Moreover, construct and incremental validity analyses confirmed that PDTQ scores relate meaningfully to some of the established correlates of misinformation susceptibility (reviewed in section 4) and predict relevant behavioural outcomes beyond existing measures (reviewed in section 5). Taken together, these results position the PDTQ as a psychometrically robust, multilingual instrument for studying perceived trust in deepfakes across diverse contexts. The final English version of the scale can be seen in Table 1.

Table 1: English version of the Perceived Deepfake Trustworthiness Questionnaire (PDTQ)

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
1.The presented information seemed convincing.	1	2	3	4	5	6	7
2.The mouth movements of the person in the video did not completely match the sound.	1	2	3	4	5	6	7
3.The background in the video contained irrelevant or out-of-place objects.	1	2	3	4	5	6	7
4.The presented information seemed plausible.	1	2	3	4	5	6	7
5.I found the voice of the person in the video unnatural.	1	2	3	4	5	6	7
6.I found the voice of the person in the video to be different from their usual voice.	1	2	3	4	5	6	7

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
7.The audio was low quality.	1	2	3	4	5	6	7
8.The presented information was something that I already know to be true.	1	2	3	4	5	6	7
9.The source of the video is verified in some way.	1	2	3	4	5	6	7
10.The facial features of the person in the video changed during the video.	1	2	3	4	5	6	7
11.The person's gestures in the video did not seem natural.	1	2	3	4	5	6	7
12.The video quality was inconsistent.	1	2	3	4	5	6	7
13.The source of the video is well-known.	1	2	3	4	5	6	7
14.The face of the person in the video (or parts of it) was distorted.	1	2	3	4	5	6	7
15.The presented information was consistent with my previous knowledge.	1	2	3	4	5	6	7
16.The source of the video seems credible.	1	2	3	4	5	6	7
17.The mouth of the person in the video was moving strangely.	1	2	3	4	5	6	7
18.The presented information seemed questionable.	1	2	3	4	5	6	7
19.The face of the person in the video (or parts of it) was blurry.	1	2	3	4	5	6	7
20.The content of the video is consistent with what this source has published previously.	1	2	3	4	5	6	7

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
21.The video was posted by a reputable source.	1	2	3	4	5	6	7
22.The presented information seemed credible.	1	2	3	4	5	6	7

Instructions: The following questionnaire contains items that aim to capture your perception of the video you just watched. Please read each item carefully and indicate your agreement using a 7-point scale ranging from »Strongly disagree« to »Strongly agree«. If you feel that you cannot answer a particular item, please choose »Neutral«.

Scoring key (R denotes that the item needs to be reverse-coded): Trustworthiness of content = (I1+I4+I8+I9+ I13+I15+I16+I18R+I20+I21+I22)/11. Trustworthiness of presentation = (I2R+I3R+I5R+I6R+I7R+I10R+ I11R+I12R+I14R+I17R+I19R)/11.

4 Beyond the Video: How Individual Differences Shape Deepfake Perception

While individuals' perception of deepfakes is a good starting point, any answers to why people believe deepfakes are incomplete without taking into account individual differences. In other words, perceived trustworthiness of deepfakes does not exist in a vacuum; instead, as demonstrated by the fact that the same videos can be perceived vastly differently by different individuals, our perception of videos is heavily influenced by our past experiences (i.e., sociodemographic variables), worldviews (i.e., motivational variables), and knowledge (i.e., cognitive variables). These factors have previously been extensively investigated in the broader misinformation context, whereas research on how they operate in the context of deepfakes and how they are specifically associated with each of the two dimensions of perceived deepfake trustworthiness is only beginning to emerge.

Starting with sociodemographic variables, previous literature has revealed that age and social media use may be important in the context of misinformation (van der Linden, 2022). In our studies, age was significantly positively associated with individuals' judgments regarding the trustworthiness of deepfakes, their content, and presentation. In other words, older individuals were more inclined to trust manipulated videos (Plohl et al., 2024). On the other hand, the frequency of using social media as a source of news was positively associated with the perceived trustworthiness of content but not the perceived trustworthiness of presentation (Plohl et al., 2024), meaning that repeated social media use may make individuals more vulnerable to questionable arguments, but may not be related to their ability to discern authentic video presentations from the manipulated ones.

Based on various theories, such as the theory of motivated reasoning, which explains that decisions are often based on pre-determined goals and desirability rather than an accurate reflection of the evidence (Kunda, 1990), researchers have identified a few individual variables that may motivate the person to believe misinformation they are exposed to. These include political orientation (Chen et al., 2023; van der Linden, 2022), belief in conspiracy theories, and trust in institutions such as media, when the media at hand is not reliable (Chen et al., 2023). Our study (Plohl et al., 2024) suggests that the importance of these factors translates to the deepfake context to some degree, but that there is an additional complexity to judging deepfakes due to their multimodal nature. Specifically, conservatism was positively associated with the perceived trustworthiness of deepfake content but was not associated with the perceived trustworthiness of presentation at all, demonstrating informational bias but no difference in deepfake recognition skills pertaining to their presentation and technical aspects.

Additionally, our unpublished results, obtained during the validation study, showed no association between conspiracy beliefs and the two dimensions measuring the perceived trustworthiness of deepfakes. As such, the role of conspiracy mentality in the perception of deepfakes remains relatively unclear. It is likely that this variable is highly context-specific; in general, it may increase distrust in the presented information, however, when deepfakes advocate for conspiracy theories, it may increase perceived trustworthiness. Lastly, in our study, trust in media was significantly positively associated with the perceived trustworthiness of deepfake content but not the perceived trustworthiness of deepfake presentation. It hence seems likely that trust in media represents a double-edged sword; trust is a necessary ingredient in communication, facilitating the spread of credible information, but, when unwarranted, it may make individuals more vulnerable to deception – a phenomenon known as misplaced trust (O''Brien et al., 2021).

In addition to demographic and motivational variables, previous research has also explored the role of cognitive abilities and other related variables. The so-called inattention account posits that being bombarded with information, coupled with limited time and resources, interferes with individuals' ability to accurately reflect on the content (van der Linden, 2022). In line with this, previous research has found that education, media literacy, reflective thinking (i.e., ability to suppress intuition and cognitively reflect when making decisions; Frederick, 2005), and so-called "bullshit receptivity" (i.e., ascribing profundity to randomly generated sentences;

Pennycook & Rand, 2019) are relatively consistently associated with the processing of misinformation, even when the content is congruent with individuals' pre-existing beliefs (Roozenbeek et al., 2020; van der Linden, 2022). In our study, we found that education was not significantly associated with the perceived trustworthiness of deepfake content or presentation. In contrast, we found significant associations between media literacy, reflectiveness, and "bullshit receptivity" on one side and the trustworthiness of content on the other side, with "bullshit receptivity" emerging as a particularly strong contributing factor. However, none of these cognitive variables were significantly associated with the trustworthiness of the presentation. The only cognitive variable significantly (albeit weakly) related to the perceived trustworthiness of presentation, not just content, was specific deepfake knowledge (Plohl et al., 2024). This suggests that while general cognitive tendencies shape how individuals evaluate the credibility of content, knowledge specific to deepfakes plays a uniquely important role in shaping perceptions of their presentation.

Table 2: A summary of factors associated with perceived trustworthiness

Category	Potential factor	Perceived trustworthiness of content	Perceived trustworthiness of presentation
Demographic variables	Higher age	✓ (↑ Risk)	✓ (↑ Risk)
	Higher social media use	✓ (↑ Risk)	X
Motivational variables	Higher political conservatism	✓ (↑ Risk)	X
	Higher belief in conspiracy theories	X	X
	Higher trust in media	✓ (↑ Risk)	X
Cognitive variables	Higher education	X	X
	Higher media literacy	✓ (↓ Risk)	X
	Higher reflective thinking	✓ (↓ Risk)	X
	Higher "bullshit receptivity"	✓ (↑ Risk)	X
	Higher deepfake knowledge	✓ (↑ Risk)	✓ (↓ Risk)

Source: Plohl et al. (2024).

As shown in Table 2, our results suggest that many known correlates of misinformation susceptibility are also relevant in the context of deepfakes. In line with this, deepfakes may disproportionally affect older individuals who use social media to a greater extent, are more conservative, trust (media) to a higher degree, have lower media literacy, are less reflective, and are more receptive to finding meaning in pseudo-profound information. The use of our scale offers additional insights. While more studies are needed, most of these factors are consistently associated with individuals' perception of the messages conveyed in deepfakes but not so much with their perception of deepfakes' presentation, which includes paying attention to the person in the video and technical aspects. In fact, only age (risk factor) and deepfake knowledge (protective factor) were associated with the perceived trustworthiness of deepfakes' presentation.

5 When Trust Turns into Influence: The Role of Perceived Trustworthiness in Shaping Attitudes and Intentions

In the previous sections, we established that people are generally bad at detecting deepfakes and provided some insight into why this is so (i.e., due to their perceptions of content and presentation, as well as demographic, motivational, and cognitive individual differences). As we approach the end of the chapter, it is worth noting why low detection and, specifically, perceived trustworthiness of deepfakes matter beyond just providing a better understanding of individuals' perception of deepfakes. We will specifically focus on associations with attitudes (i.e., psychological tendencies expressed by evaluating a particular entity with some degree of favour or disfavour; Eagly & Chaiken, 1993) and behavioural intentions (i.e., individuals' intention to perform a given act; Ajzen & Fishbein, 1972) - two outcomes related to behaviour (Ajzen, 1991).

One of our studies showed that low detection, across various deepfake videos, led to more favourable affective responses to videos (i.e., higher liking), which, in turn, led to increased intentions to share the manipulated videos on social media (Plohl et al., 2025b). Similar associations were found between sharing intentions and perceived trustworthiness of deepfakes, with these results offering additional insight into the complex relationship between variables. Specifically, in the original PDTQ validation study (Plohl et al., 2024), we investigated whether perceived trustworthiness of content and presentation explain variance in viral behavioural intentions (i.e., the intentions to like, share, and recommend the video) beyond basic demographic variables (i.e., age, education, political conservatism, social media use), individual differences (i.e., "bullshit receptivity", reflectiveness, trust in media, media literacy, deepfake knowledge), and a previous scale measuring participants' perception of the manipulated video (i.e., Message Believability Scale; Hameleers et al., 2023). We found that the newly developed scale explained a significant part of the variance (an additional 5.0%) in viral behavioural intentions over and above other included variables. In the final model, which was able to explain 36.0% of the variance, age, "bullshit receptivity", reflectiveness, trust in media, deepfake knowledge, message believability, and trustworthiness of content, which was the strongest predictor, significantly predicted the outcome. Other variables, including the trustworthiness of the presentation, did not significantly predict viral behavioural intentions. These results suggest that individuals' intention to spread the videos may be particularly driven by the trustworthiness of the content. Nonetheless, the questionnaire explained a significant additional share of variance, highlighting the added value of a more comprehensive measurement of deepfake perception.

The importance of these perceptions was further demonstrated in our experimental study (Plohl et al., 2025a), which examined the potential positive or negative effects of a single exposure to deepfake or authentic videos on individuals' attitudes toward climate change and immigration, two highly polarized, politically sensitive issues (Doss et al., 2022; Hameleers et al., 2022; Westerlund, 2019). Specifically, the study explored boundary conditions under which attitude change might occur, with a focus on video quality, perceived trustworthiness, and political alignment.

A total of 1,124 participants from the United Kingdom, Italy, and Slovenia watched real videos, high-quality deepfakes, or low-quality deepfakes advocating for or against climate action and immigration (Figure 1). The quality of videos was manipulated in terms of the content and presentation. For example, manipulations of content included changing the supposed source of the video and making the presented information more or less aligned with the target person's actual stance on the topic. In contrast, manipulations of presentation included alterations of mouth movements, voice, and video quality. All videos lasted approximately one minute and featured well-known proponents or opponents of climate change and immigration. Participants provided their demographic data and filled out the PDTQ (Plohl et al., 2024) directly after watching each of the two videos, whereas the Scepticism scale (a measure of attitudes towards climate change; Whitmarsh, 2011) and the Positive and Negative Perception of Immigrants Scale (a measure of attitudes towards immigration; Panno et al., 2023) were filled out before and after video exposure.

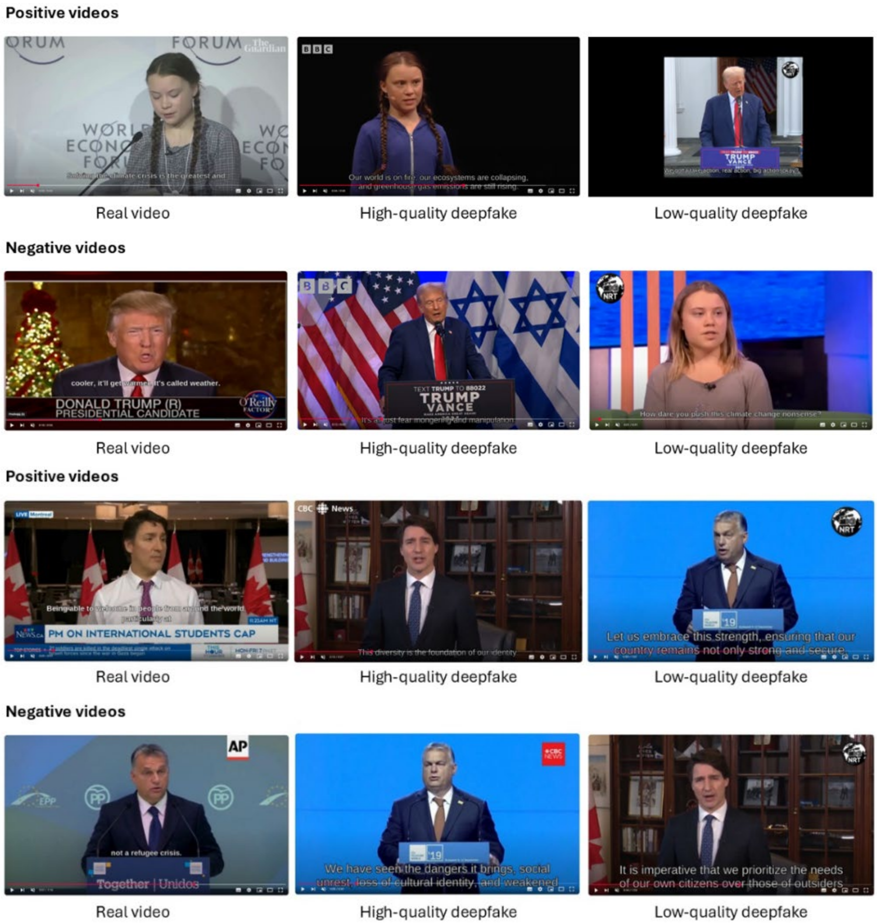

Positive videos
Real video
High-quality deepfake
Low-quality deepfake
Negative videos
Real video
High-quality deepfake
Low-quality deepfake
Positive videos
Real video
High-quality deepfake
Low-quality deepfake
Negative videos
Real video
High-quality deepfake
Low-quality deepfake
Our world is on fire, our ecosystems are collapsing, and greenhouse gas emissions are still rising
cooler, it'll get warmer. It's called weather.
DONALD TRUMP (R) PRESIDENTIAL CANDIDATE
How dare you push this climate change nonsense?
PM ON INTERNATIONAL STUDENTS CAP
This diversity is the foundation of our identity
Let us embrace this strength, ensuring that our country remains not only strong and secure
not a refugee crisis.
Together | Unidos
We have seen the dangers it brings, social unrest, loss of cultural identity, and weakened
It is imperative that we prioritize the needs of our own citizens over those of outsiders

Figure 1: Stimuli related to climate action (first two rows) and immigration (last two rows)
Source: own.

Contrary to expectations, neither video authenticity/quality nor political orientation moderated the impact of the videos on attitudes. On the other hand, perceived trustworthiness of deepfake content consistently predicted attitude change across both topics, while perceived presentation trustworthiness was associated with attitude shifts on immigration. Specifically, when individuals watched a video emphasizing that climate change is real and promoting positive attitudes towards immigrants and perceived it as highly trustworthy in terms of the content, this perception had larger positive effects on attitudes (and vice versa for videos opposing climate change and communicating negative attitudes towards immigrants). Similarly, when individuals perceived the immigration video as highly trustworthy in terms of the presentation, the videos emphasizing positive attitudes towards immigrants exhibited larger positive effects on attitudes (and vice versa for videos communicating negative attitudes towards immigrants). These findings indicate that subjective perceptions of trustworthiness, rather than objective video features or ideological congruence, are central to understanding how deepfakes shape public opinion. Interestingly, our results also suggest that the perceived trustworthiness of a video's content exerts a more consistent and stronger effect than its presentation. Although visual and technical elements can enhance a video's sense of realism, it is the plausibility and coherence of the message that seem to play the more decisive role in shaping attitudes, at least in the political sphere, where audiences often possess prior knowledge about public figures; messages that align with these expectations may be perceived as more credible, even when their presentation is less polished.

6 Concluding Remarks

In conclusion, the evidence reviewed in this chapter paints a comprehensive picture of why people believe deepfakes and how such beliefs can shape attitudes and behavioural intentions. We began by highlighting that, despite public confidence in detection abilities, people are generally poor at distinguishing deepfakes from authentic videos, often performing only slightly above chance.

We then introduced the concept of perceived trustworthiness as a way to move beyond binary detection measures and capture the perceptual factors that drive belief in deepfakes. Our work distinguishes between the trustworthiness of a video's content (i.e., how plausible and credible the message appears) and its presentation (i.e., how authentic the visual, auditory, and behavioural cues seem). This distinction reveals that, due to their multimodal nature, judgments of deepfake videos go far beyond evaluations related to the factual accuracy of the content. While both dimensions matter, trustworthiness of content emerges as more strongly linked to individual differences such as political orientation, trust in media, and cognitive reflection, and more predictive of attitudinal outcomes, perhaps because audiences are not (yet) adept at scrutinizing subtle visual or behavioural inconsistencies.

We further examined how individual characteristics spanning demographic, motivational, and cognitive factors interact with perceptual processes to shape susceptibility. Factors such as age, social media use, media literacy, "bullshit receptivity", and deepfake-specific knowledge influence whether viewers are more or less likely to accept deepfakes as genuine. Importantly, these variables are often more strongly associated with content-related trustworthiness than presentation-related trustworthiness.

Finally, we showed that perceptions of trustworthiness do not remain at the level of passive judgments; they can translate into measurable attitude change and behavioural intentions such as sharing content on social media. In our studies, the perceived trustworthiness of content consistently predicted shifts in views on polarized issues like climate change and immigration, regardless of objective video quality or political alignment. This highlights the broader persuasive potential of deepfakes; even imperfect manipulations can influence public opinion when their message resonates.

Taken together, these findings demonstrate that it is not the objective properties of a video, but the perceived credibility of its message and presentation, that drive its psychological impact. If deepfakes are a technological challenge, belief in deepfakes is a psychological one. Protecting the public will therefore require both technological detection tools and psychological interventions that address the perceptual, cognitive, and motivational factors underlying belief. In an era where seeing is no longer believing, this dual approach is essential for preserving informed decision-making, public trust, and democratic stability.

Building on this, the construct of perceived trustworthiness, along with the developed questionnaire, which represent the chapter's most significant contributions, may also guide policy and platform responses, explored in more detail in Chapter 8. Because the PDTQ quantifies how believable a deepfake appears to ordinary viewers, it can be used as an input for automated moderation pipelines or risk assessment systems, for example, by assigning each video a "harm score". Content that scores high on trustworthiness but is identified as synthetic could be prioritized for rapid review or removal, while lower-scoring deepfakes might be flagged for further verification without immediate action. Similarly, PDTQ items may be used to develop specific interventions prior to exposure and deliberation prompts at the point of exposure, helping users critically evaluate manipulative content before it shapes their beliefs or behaviour. In this way, psychological insights into why people believe deepfakes can directly inform scalable, evidence-based, and, perhaps most importantly, citizen-empowering policy responses, bridging the gap between individual-level perception and systemic prevention strategies.

End notes

All authors helped conceptualize the chapter and actively contributed to psychological studies carried out within the SOLARIS project, which are presented in the chapter. Nejc Plohl prepared the original draft, while Urška Smrke, Letizia Aquilino, and Izidor Mlakar contributed to reviewing and editing the chapter. Izidor Mlakar led this part of the project and supervised the writing process.

References

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 behavioral intentions. *Journal of Personality and Social Psychology, 21*(1), 1–9. https://doi.org/10.1037/h0031930

Chen, S., Xiao, L., & Kumar, A. (2023). Spread of misinformation on social media: What contributes to it and how to combat it. *Computers in Human Behavior, 141*, 107643. https://doi.org/10.1016/j.chb.2022.107643

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

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., Bush, L., & Tucker, C. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports, 13*(1), 13429. https://doi.org/10.1038/s41598-023-39944-3

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods, 29*(3), 494–518. https://doi.org/10.1037/met0000540

Hameleers, M., van der Meer, T. G., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media + Society, 8*(3), 20563051221116346. https://doi.org/10.1177/20563051221116346

Hameleers, M., van der Meer, T. G., & Dobber, T. (2023). They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication, 39*(1), 56–70. https://doi.org/10.1177/02673231231184703

Hameleers, M., van der Meer, T. G., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior, 152*, 108096. https://doi.org/10.1016/j.chb.2023.108096

Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 188–193. https://doi.org/10.1089/cyber.2020.0174

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience, 24*(11), Article 103364. https://doi.org/10.1016/j.isci.2021.103364

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

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

Lee, J., & Shin, S. Y. (2022). Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology, 25*(4), 531–546. https://doi.org/10.1080/15213269.2021.2007489

O''Brien, T. C., Palmer, R., & Albarracin, D. (2021). Misplaced trust: When trust in science fosters belief in pseudoscience and the benefits of critical evaluation. *Journal of Experimental Social Psychology, 96*, 104184. https://doi.org/10.1016/j.jesp.2021.104184

Panno, A., Pellegrini, V., De Cristofaro, V., & Donati, M. A. (2023). A measure of positive and negative perception of migration: Development and psychometric properties of the Positive and Negative Perception of Immigrants Scale (PANPIS). *Analyses of Social Issues and Public Policy, 23*(1), 73–105. https://doi.org/10.1111/asap.12338

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality, 88*(2), 185–200. https://doi.org/10.1111/jopy.12476

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

Plohl, N., Mlakar, I., Aquilino, L., Brienza, M., Bisconti, P., & Smrke, U. (2025a). The moderating role of perceived trustworthiness in explaining the attitudinal effects of political deepfakes. *SSRN Preprint.*https://doi.org/10.2139/ssrn.5351533

Plohl, N., Mlakar, I., Aquilino, L., Brienza, M., Bisconti, P., & Smrke, U. (2025b). How deepfake quality, media literacy, and personal attitudes shape detection, liking, and social media sharing of political deepfakes. *PsyArXiv Preprint.*https://doi.org/10.31234/osf.io/knvby

Plohl, N., Mlakar, I., & Kecelj, Ž. (2025c). *Investigating the effects of an educational infographic and cognitive load on deepfake detection and metacognitive judgments.* Manuscript in preparation.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science, 7*(10), 201199. https://doi.org/10.1098/rsos.201199

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior, 149*, 107917. https://doi.org/10.1016/j.chb.2023.107917

Somoray, K., Miller, D. J., & Holmes, M. (2025). Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies, 2025*(1), Article 1833228. https://doi.org/10.1155/hbe2/1833228

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M. A., & Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445699

Thaw, N. N., July, T., Wai, A. N., Goh, D. H. L., & Chua, A. Y. (2021). How are deepfake videos detected? An initial user study. In *Proceedings of the 23rd HCI International Conference, HCII 2021, Part 1* (pp. 631–636). Springer. https://doi.org/10.1007/978-3-030-78635-9_80

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine, 28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53. https://doi.org/10.22215/timreview/1282

Whitmarsh, L. (2011). Scepticism and uncertainty about climate change: Dimensions, determinants and change over time. *Global Environmental Change, 21*(2), 690–700. https://doi.org/10.1016/j.gloenvcha.2011.01.016

DOI
https://doi.org/
10.18690/um.feri.2.2026.5

ISBN
978-961-299-109-8

Democracy Distorted – Deepfakes as Political Weapons

Gjon Rakipi,[1] Yasaman Yousefi,[2, 3]
Calogero Caltagirone,[4] Angelo Tumminelli,[4]
Andrew McIntyre,[5] Aseniya Dimitrova[6]

[1] Albanian Institute for International Studies (AIIS), Tirana, Albania
gjonrakipi@aiis-albania.org
[2] DEXAI-Artificial Ethics, Rome, Italy
yasaman.yousefi@dexai.eu
[3] University of Bologna, CIRSFID ALMA AI, Faculty of Legal Studies, Bologna, Italy
y.yousefi@unibo.it
[4] LUMSA University, Department of Human Sciences, Rome, Italy
a.tumminelli@lumsa.it, c.caltagirone@lumsa.it
[5] University of Amsterdam, Institute for Logic, Language and Computation; Amsterdam, the Netherlands
a.mcintyre@uva.nl
[6] Brand Media Bulgaria, Sofia, Bulgaria
seniya.dimitrova@gmail.com

Affordable generative AI allows actors to produce and amplify deepfakes instantly, outpacing verification efforts. Drawing on Young's (2011) distinction between isolated harms and structural injustice, this chapter identifies synthetic media as a structural threat to democracy that collapses the evidentiary foundations of public reason. We examine how deepfakes weaponize information ecosystems, using European and U.S. case studies to demonstrate their specific deployment against women and minority candidates. Methodologically, we analyse recent disinformation incidents through the lenses of epistemic injustice and deliberative democracy. We argue that deepfakes signal a deeper vulnerability where truth becomes malleable and public trust erodes. The chapter concludes that safeguarding democratic life requires not only legal and technical fixes, but a normative reorientation toward truthfulness and accountability.

Keywords:
disinformation,
electoral integrity,
epistemic injustice,
gendered harms,
democratic resilience

University of Maribor Press

1 Conceptualising Harm

Harm is an elastic idea. In its oldest sense, it names any blow to a person's well-being: a broken bone, a stolen wage, a silenced voice. Yet the digital century invites a broader lens. Today, a manipulated recording, such as the AI-generated audio targeting Michal Šimečka just days before Slovakia's 2023 vote (Meaker, 2023), can circulate in the morning, fracture public trust by noon, and tilt an election by evening. Such episodes remind us that harm is both material, and epistemic and political. Epistemic harm occurs when the channels through which we come to know the world are deliberately muddied. Deepfakes, coordinated rumour campaigns, and AI-generated "news" flood the evidentiary pool with noise, making it harder for individuals to sort fact from fabrication. Uncertainty is not a neutral by-product here; it is the intended wound, eroding a community's capacity to share reasons and reach common judgments. Political harm builds on this erosion. Democratic life depends on citizens who can verify, contest, and ultimately consent to the decisions made in their name. When falsehoods travel faster than rebuttals, accountability mechanisms falter. The result is not just misinformed voters but a weakening of the very norms that make collective self-government possible. By foregrounding these layered harms, the chapter can shift from cataloguing threats to explaining why they matter normatively, providing the conceptual framework we will use to analyse gendered disinformation (Section 5.4) and the erosion of democratic values (Section 5.5). Readers will see that the stakes extend beyond isolated victims to the cognitive and institutional scaffolding on which democratic societies rest.

2 Electoral Interference in Europe and Beyond

Elections are pivotal moments for democratic societies, where this single event can significantly alter power structures, policy directions, and political representation at local, national, and international levels. Both are the outcomes of elections highly consequential, but they also often trigger periods of intense political engagement and polarization among citizens, as competing socio-political messages come to the forefront of public discourse and debate. Additionally, elections are highly mediated events as political parties, and their supporters communicate their messages to the public via a wide range of media channels (Mazzoleni & Schulz 1999). This includes campaign materials (e.g., posters, adverts, leaflets), political activities (e.g., speeches, press conferences) and journalistic coverage (e.g., opinion pieces, interviews, televised debates). This mediatization of elections has only intensified with the rise of social media platforms, wherein political content can be directly communicated to individual users in a highly personalized way through network connections, algorithmic recommendations, and targeted advertising (Marwick & Lewis 2017; Chun 2021).

The combination of highly consequential outcomes, a politically sensitive environment, and the pervasive mediation of political messaging means that elections are particularly attractive and vulnerable targets for political manipulation through coordinated disinformation campaigns. Given these factors, even the uncoordinated and/or unintentional spread of disinformation during election periods can have a significant impact.

With the arrival of modern generative AI systems and the widespread production and spread of synthetic media online, elections have become ever more dangerous times for democratic societies. Generative AI systems are now capable of producing high-quality synthetic audiovisual content (e.g., images, video, audio, text) that is near-indistinguishable from authentic content (Yazdani et al. 2025). Furthermore, the arrival of these systems means that the production of high-quality disinformation is less costly (Smith and Mansted 2020). Synthetic media depicting government officials, political figures and influential media personalities doing or saying anything could have a significant impact on the outcome of elections (Chesney & Citron, 2019; Diakopoulos & Johnson, 2019). For example, such content could be used to undermine the reputation of public figures, deceptively sway public opinion on specific issues, and/or threaten influential figures to manipulate their actions and political positions.

Since the emergence of deepfakes in 2017, there have already been numerous high-profile cases of synthetic media being used for electoral interference. For example, in the run up to the Slovak parliamentary elections in 2023, synthetic audio released online appeared to show politician Michal Šimečka, leader of the Progressive Slovakia party, discussing plans to rig the election in an attempt to undermine his credibility in the eyes of voters (Meaker, 2023). Meanwhile, the 2024 Pakistan general election saw several synthetic audiovisual recordings circulating online. These appeared to show prominent members of the Pakistan Tehreek-e-Insaf (PTI) party, including imprisoned leader Imran Khan, calling for a boycott of the election meant to deceive PTI supporters into abstaining (Tiwari, 2024). In both cases, the synthetic content was identified as inauthentic by news media and the impact upon the election was seemingly minimal. Progressive Slovakia came second in the parliamentary elections, while in Pakistan PTI-backed candidates won more seats than any other single party. Ironically enough, Khan declared victory from jail using synthetic media. Though technically convincing, synthetic content that misrepresents high-profile political figures like Šimečka and Khan is unlikely to deceive a significant proportion of the public to have a considerable impact. This is because such content receives considerable attention and scrutiny to be easily detected and debunked. What is less widely discussed, but potentially more dangerous to electoral integrity, is the use of synthetic media in low-profile political settings; so-called "microfakes".

Where high-profile disinformation is likely to be debunked, synthetic content depicting figures and officials involved in smaller-scale politics may go undetected as such content is unlikely to be widely distributed and properly scrutinized (Ascott, 2020). Smaller-scale disinformation campaigns featuring local politicians or officials addressing local controversies (e.g., road quality, bypass development, cycle lanes) may appear technically convincing and interfere with local elections. Though there is currently little evidence of real-world microfakes, cases are unlikely to be reported by their very nature. As one clear example, during the 2022 mayoral election in Shreveport, Louisiana, the likeness of incumbent Democratic candidate Adrian Perkins was digitally recreated using AI as part of a hostile political advertisement criticising his policies (Swenson et al. 2024). Perkins ultimately lost the election and claims this deepfake advertisement played a crucial role. Though openly artificial and intended as humorous satire, this advertisement proves that such microfakes could be utilized at a local level. While the immediate impact of these microfakes may be minor, coordinated disinformation campaigns targeting numerous local elections could represent a granular and gradual threat to democracy that escalates to influence national and international politics.

Beyond disinformation campaigns aimed directly at undermining the credibility of candidates or influencing voter sentiments on specific issues, synthetic media can also be used to intimidate, threaten or otherwise harass political figures to influence their actions and statements, or to deter political participation altogether (Chesney & Citron, 2018). Notably, the production of deepfake pornographic content presents a significant reputational risk and thus the very threat of publication could be used to deter candidates from standing in elections, as will be discussed in more detail below (Adjer et al., 2019; Rini & Cohen, 2022).

While the arrival of generative AI may be exacerbating risks for electoral interference, synthetic content emerged into an information environment that was already fertile ground for rampant disinformation and post-truth politics. Throughout the 2010s and into the 2020s, there has been a noted decline in traditional news media as people have grown more dependent on social media platforms as the primary source of political information. Unlike traditional journalism which relies on editorial standards and fact-checking, social media platforms operate and disseminate content according to an attention economy wherein there is such an overabundance of content that the flow of information hinges upon what will attract people's attention immediately (Lewis & Marwick, 2017). Such a system prioritizes emotionally charged or sensational content rather than complex, nuanced information. More so than traditional media. As such, these networks allow for disinformation and false narratives to circulate widely among platform users before traditional journalists and fact-checkers can publish evidence-based rebuttals or corrections. Within this attention economy, sensational political synthetic media may spread online too rapidly or go entirely unnoticed, potentially influencing users that have little media literacy skills or that are less engaged with broader political discourse and debates. These networks are also extremely vulnerable to attention-hacking techniques that seek to manipulate those content filtering and recommendation algorithms that dictate what information users see and interact with. For example, throughout the 2010s, far-right extremists frequently coordinated large groups of users to flood Twitter with specific hashtags (e.g., #gamergate) to artificially make this topic trend and reach users who might not otherwise encounter their propaganda. In other instances, these extremists have piggybacked on already trending hashtags (e.g., #blacklivesmatter) to hijack its popularity and strategically amplify the reach of their own political messages.

Designed to capitalise on this attention economy, algorithmic recommendation systems preferentially show users content that provokes engagement. In doing so, these systems reinforce pre-existing biases and deepen divisions along ideological lines. Building on this algorithmic polarization, users of online platforms are increasingly connected based on the principle of homophily i.e., the assumption that similarity breeds connection (Chun, 2024). Algorithmic recommendation systems cluster individual users into neighbourhoods based on similarity (e.g., race, gender, sexuality, political affiliation). This clustering encourages political echo chambers to form wherein there is little exposure to conflicting information and people are encouraged to accept information that confirms their existing beliefs, regardless of its accuracy. Within such neighbourhoods, political messaging and disinformation can spread freely and with greater impact via strong interpersonal ties among members. Synthetic content promoting false political narratives can, therefore, be more readily accessed, accepted and shared. Once embedded, these false narratives are difficult to combat, shaping voter perceptions and undermining trust in the legitimacy of democratic societies.

More generally, the proliferation of synthetic media that is near-indistinguishable from authentic content means that people are more sceptical of all information they receive online, and they are less likely to trust traditional information sources and authorities (Vaccari & Chadwick, 2020). The epistemic impact of synthetic media on our information environment more broadly is discussed in the next section.

3 Epistemic Erosion and the Misinformation Ecosystem

Beyond headline elections, deepfakes exacerbate the chronic "liar's dividend": the mere possibility that any footage might be fabricated empowers bad actors to dismiss authentic evidence and fuels public cynicism. A 2024 European Parliamentary briefing warns that synthetic media risks a downward spiral in which voters "no longer believe what they see or hear," undermining media pluralism and parliamentary scrutiny (Michael & Hocquard, 2023). UNESCO's report (2023) on freedom of expression during elections similarly notes that cheap-fakes and deepfakes erode basic informational rights by diffusing responsibility among anonymous creators, automated recommender systems, and inattentive platforms. Experimental work published in Digital Journalism finds that high-quality deepfakes reduce viewers' trust in both the target and the outlet that hosts the correction, even when the fabrication is revealed within seconds (Patel, 2025). The study referenced, published in the journal Digital Journalism, is part of a growing body of research examining the impact of deepfakes on public trust. Deepfakes are AI-generated manipulated videos capable of producing extremely realistic footage, often difficult to distinguish from authentic content. The researchers conducted controlled experiments in which participants were shown short, high-quality deepfake videos, followed by an immediate correction or debunk published by a news outlet. The interval between viewing the deepfake and being informed of its falsity was only a few seconds, an intentionally "ideal" scenario in which both the victim and the news organization respond as quickly and transparently as possible. The cumulative outcome is an epistemic environment where strategic actors can manufacture plausible doubt faster than institutions can generate consensus, eroding the public's capacity for informed deliberation.

3.1 Infodemic and Epistemic Erosion: The Role of Deepfakes

An infodemic is a phenomenon in which an excessive amount of unverified or contradictory information makes it difficult for recipients to ground themselves in reality (World Health Organization, 2020; Cinelli et al., 2020). The category of "infodemic" has gained importance, especially during the COVID-19 pandemic, but it represents a broader and ongoing issue that is linked to the digital age in which news, true, false, or distorted, spreads at unprecedented speeds.

This is the context in which a subset of generative AI known as deepfakes emerges. Deepfakes are able to bolster the infodemic, making it increasingly difficult to distinguish between what is authentic and what is manipulated. Their impact is both informative and epistemic in that they undermine our ability to trust traditional sources and media, reconfiguring the very modalities of knowledge and perception of the world.

This epistemic erosion weakens the pact of trust on which shared knowledge is based. In fact, when even digital content can be manipulated in a dystopian way, our perception of reality itself becomes fragile and fuels an informational relativism that opens the doors to a dangerous revisionism and systemic distrust.

Without critical tools and adequate regulatory frameworks, we risk having a society in which the truth is not only manipulable but also completely delegitimized. To counter this drift, it is necessary to invest in media literacy and accountability.

AI certainly represents one of the most insidious challenges for public information in the 21st century: it is a non-neutral tool that, if used maliciously, can become a powerful vehicle for disinformation and epistemic dystopia. In fact, in public contexts, such as politics, journalism, or social debate, deepfakes undermine the reliability of content and contribute to eroding truth as the foundation of collective discourse (Weikmann & Lecheler, 2024). This determines the phenomenon that has been appropriately defined as "epistemic pollution" with which information is distorted, manipulated or presented in a misleading way, compromising our ability to know and understand the world (Levy, 2021). In a dystopian context, the use of artificial intelligence can amplify this phenomenon, generating intentionally false but credible content. Algorithms trained on partial or manipulated data can reinforce pre-existing biases, creating information bubbles and cognitive polarization (Praiser, 2011; O''Neil, 2016). This phenomenon fuels a dangerous form of information nihilism (Labarre, 2025), in which every truth is suspect, every piece of evidence is revocable, and every opinion becomes equally valid. In such a climate, truth loses its value and illusion takes over. The consequences are profound: social polarization, civic disillusionment, and the delegitimization of democracy. Furthermore, and very relevant to this reflection, AI can be used by authoritarian regimes or interest groups to rewrite historical and cultural narratives (Hameleers et al., 2024). In the absence of transparency and control, reliable sources lose relevance, and access to knowledge is filtered by opaque interests. Information democracy turns into an algorithmic oligarchy that must be countered through critical awareness and the ethical governance of AI.

Addressing the impact of deepfakes requires rethinking verification standards, promoting digital literacy, and holding content creators and platforms accountable. Only through these efforts can truth be defended in an increasingly vulnerable public sphere.

3.2 The ethical dimensions of deepfakes

Deepfakes blur the line between authentic and fabricated evidence, threatening individual autonomy and public trust. This has serious implications for fields like journalism and law enforcement, where visual evidence plays a critical role. Fabricated content in these areas can have far-reaching consequences, including the corruption of the historical record, the miscarriage of justice, and the undermining of public trust in essential institutions. The issue of consent is also paramount when it comes to deepfakes. Using someone's likeness without their agreement, particularly for harmful purposes, violates personal rights and dignity. The potential use of deepfakes in international relations adds another layer of complexity to the ethical debate. They could be used to create false evidence, to mislead the public or international community, and potentially to provoke conflicts or exacerbate

4 Gendered and Minority Harms

As discussed in earlier sections, the advent and diffusion of synthetic media technologies, particularly deepfakes, pose significant challenges to democratic life. However, it is essential to recognize that these harms are not borne equally. An emerging body of evidence demonstrates that the impacts of deepfakes are disproportionately experienced by women and minority groups, both in their private lives and in the public sphere. This section examines how deepfakes operate as technological amplifiers of entrenched social inequalities, drawing on empirical research, legal scholarship, and documented case studies to articulate their normative and political consequences.

A pivotal moment in this discourse came with the 2019 audit conducted by the cybersecurity firm Deeptrace. Their findings revealed that 96 percent of the 14,678 deepfake videos indexed at that time were non-consensual pornographic content targeting women (Adjer et al., 2019). Subsequent studies have since corroborated this troubling trend. For instance, a 2024 survey spanning ten countries found that 2.2 percent of respondents reported being targeted by synthetic intimate imagery without their consent, with women and gender minorities disproportionately represented among the victims (Umbach et al., 2024). These figures illustrate a broader phenomenon: the weaponization of deepfake technology to perpetuate gender-based violence and harassment.

While the development of generative AI was initially confined to research circles, this changed in 2017 when a Reddit user under the pseudonym "Deepfakes" began distributing manipulated pornographic videos using free, open-source machine learning tools. This marked a turning point in the accessibility and misuse of synthetic media, setting a precedent for widespread abuse.

Academic literature has repeatedly emphasized the gendered nature of deepfake harms. Chesney and Citron have argued that non-consensual deepfake pornography, as one of the earliest and most prevalent applications of the technology, systematically targets women and introduces novel forms of gender-based abuse. With minimal technical expertise, perpetrators can now fabricate highly realistic sexual content using another person's likeness, thereby enabling a continuum of exploitative practices that includes sextortion, reputational sabotage, blackmail, and intimate partner violence (Chesney & Citron, 2018). Yet the scope of exploitation is not limited to sexualized media. Deepfakes have also been deployed in cases of identity fraud, financial scams, and emotional coercion, including fabricated kidnapping videos or synthetic recordings designed to manipulate or intimidate. These forms of abuse are not merely technological anomalies; they reflect deeper structural patterns in which individuals are rendered tools for others' gain, often at great personal and societal cost.

Laffier and Rehman have further highlighted the psychological and reputational consequences of these abuses, noting that victims frequently suffer job loss, social exclusion, and severe mental health outcomes (Laffier & Rehman, 2023). The weaponization of deepfakes against women and minority communities thus functions as a form of personal attack and as a mechanism for reinforcing existing social hierarchies and exclusions.

In political contexts, these harms have a particularly corrosive effect on democratic participation. Deepfakes increasingly operate as tools of deterrence, strategically targeting underrepresented groups to dissuade them from civic engagement. They undermine the democratic ideal of equal participation by selectively amplifying social vulnerabilities and exploiting pre-existing prejudices. Female politicians, already the subject of disproportionate online abuse, now contend with the added threat of AI-generated disinformation. Such campaigns are capable of producing fabricated pornographic material, falsified news articles, and synthetic audiovisual recordings, all designed to erode credibility and sow distrust.

One of the most troubling aspects of gendered disinformation is its adaptability. Algorithmic systems can customize fabricated content to match the biases of particular audiences (Goldstein et al., 2023). In conservative-leaning electorates, such content may depict women in line with regressive gender stereotypes, questioning their emotional stability or capacity for leadership. In more progressive regions, false narratives may be engineered to simulate scandal or ethical misconduct. Regardless of context, the end goal remains the same: to undermine a woman's professional and political legitimacy.

The deployment of deepfakes in electoral politics is increasingly well-documented. In France, ahead of the 2024 EU elections, deepfake videos circulated online purporting to show young women identified as nieces of Marine Le Pen endorsing far-right ideologies. These videos, though fabricated, gained significant traction and sparked renewed debate over the inadequacy of content moderation in responding to political disinformation (Hartmann, 2024). In Germany, during the 2021 federal election, Annalena Baerbock, the Green Party's candidate for Chancellor, was the target of AI-generated narratives laced with gendered tropes and intimidation tactics. These efforts compromised her individual campaign, and sent a chilling message to women contemplating political careers (Kovalčíková & Weiser, 2021). In Italy, female politicians across the political spectrum, including Prime Minister Giorgia Meloni and opposition leader Elly Schlein, have been targeted with deepfake pornography and sexually explicit images, forming part of a broader strategy of delegitimization through misogynistic content (Chopra et al., 2025; Giuffrida, 2025).

These attacks are part of a broader strategy of participatory deterrence. By inflating the reputational and personal costs of public life, deepfakes serve to exclude marginalized groups from democratic institutions. The concept of epistemic injustice, as theorized by Miranda Fricker, proves useful here, specifically her notion of 'testimonial injustice,' which describes how prejudice leads audiences to assign a 'credibility deficit' to a speaker, wrongly stripping them of their status as a reliable knower. (Fricker, 2007). It captures the systematic devaluation of certain groups as credible knowers and participants in public discourse. Deepfakes exacerbate such injustice by selectively targeting those who already face structural disadvantages, thereby intensifying their marginalization. The result is an informational environment in which appearances override evidence, and democratic deliberation gives way to aesthetic manipulation, echoing concerns about an emerging "post-truth geopolitics" (Chesney & Citron, 2019).

A further challenge lies in the responses, or lack thereof, by digital platforms. Social media companies and content-sharing platforms often treat pornographic deepfakes as privacy issues rather than as democratic threats. Consequently, moderation and takedown mechanisms tend to lag behind the speed at which such content spreads, allowing politically motivated synthetic media to reach wide audiences before fact-checkers can intervene (Chesney & Citron, 2018). This regulatory inertia enables malicious actors to exploit algorithmic amplification and virality, often with impunity.

The harm is amplified by the architecture of digital platforms themselves. Deepfakes can be created with basic tools, uploaded in seconds, and rapidly disseminated across networks at little to no cost. Victims and public institutions frequently struggle to keep pace. Even after content is debunked, its reputational damage often persists, illustrating the profound temporal and institutional asymmetries embedded in the current media ecosystem.

Compounding this situation is a failure of governance. Carpenter notes cheap-fakes and deepfakes fracture the informational commons by diffusing accountability across anonymous creators, automated content delivery systems, and disengaged platform policies. The result is an epistemic landscape where both truth and trust are undermined, and where the mere possibility of fabrication, the so-called "liar's dividend", is sufficient to discredit even authentic evidence (Carpenter, 2024).

In sum, the gendered and minority harms of deepfakes are not isolated incidents but structural phenomena that exploit existing inequalities, distort democratic processes, and degrade informational integrity. Addressing these harms demands, at a superficial level, technical fixes and, more profoundly, a normative reorientation that centres justice, accountability, and inclusive participation in the governance of emerging technologies.

5 Normative Implications for Democratic Values

Liberal democracy relies on citizens being able to verify what leaders say and do. When a convincing AI-generated video or audio circulates, that shared evidentiary ground can disappear. Deliberative theorists such as John Rawls describe this ground as the basis of public reason, the arena where disagreements are settled with facts that everyone can inspect. Deepfakes undermine that arena in two reinforcing ways. First, they insert persuasive falsehoods faster than journalists and fact checkers can react. Second, the very existence of generative forgeries lets dishonest actors deny authentic evidence. This forementioned liar's dividend means that someone caught in wrongdoing can claim the incriminating video is merely synthetic (Chesney & Citron, 2018). Both dynamics erode transparency because they make visual or auditory proof negotiable rather than authoritative.

The European Union's Artificial Intelligence Act (Article 50) will require clear labelling of synthetic audiovisual content to restore minimum transparency, but enforcement will not begin until the regulation's phased entry into force in 2025 (*European Union Artificial Intelligence Act: A Guide*, 2025). Until then, Europeans inhabit what philosopher Regina Rini describes as an epistemic fog where seeing is no longer believing.

Democracy promises that every citizen's contribution deserves comparable credibility. Deepfakes threaten this promise by amplifying pre-existing asymmetries of capacity and access. Producing convincing synthetic media still demands specialized skills, substantial computing power, or paid software, whereas evaluating authenticity usually requires time, digital literacy, and sometimes proprietary forensic tools. Well-resourced actors, for example, large campaigns, state broadcasters, or private influence firms, therefore, enjoy a comparative advantage in shaping narratives, while ordinary citizens must consume content in real time without equivalent verification resources. One 2019 article notes that deepfake operations concentrate communicative power in the hands of those with technical sophistication, and such a concentration is able to skew public deliberation toward elites with asymmetric informational control (Kietzmann et al., 2020).

From a deliberative perspective, the problem is not simply unequal speech volume, but unequal credibility allocation. Citizens lacking digital-forensic literacy are more likely to accept forged media as real or to dismiss genuine media as fake, creating what epistemologists describe as credibility deflation, a systemic reluctance to trust anyone who lacks signals of technological authority. Rural populations, older voters, and linguistic minorities often face additional barriers to reliable verification services, perpetuating a civic hierarchy in which those with access to advanced tools can define what counts as knowledge. Equality suffers even without targeted harassment because the communicative space tilts toward actors who can purchase sophisticated deception or rapid authentication.

Transparency failures and credibility gaps combine to weaken accountability, the process that turns democratic judgment into real consequences. Deepfakes enable false scandals to destroy reputations overnight, and let genuine misconduct be waved away as "fake" procedures meant to encourage calm reflection can be hijacked by synthetic evidence that spreads suspicion when replies are legally muted.

Jürgen Habermas stresses that democratic legitimacy rests on communicative rationality, a norm requiring actors to justify their positions with reasons subject to public testing. Deepfakes loosen the bond between action and proof, enabling officials to evade substantive answers by questioning the medium itself. The public sphere risks sliding toward post-truth politics, a climate in which empirical validation yields to partisan loyalty.

Transparency, equality, and accountability form an interlocking architecture. When transparency falters, resource-rich actors exploit the uncertainty, which deepens inequality in communicative power. That inequality then makes it easier for influential players to deploy or dismiss synthetic media, further weakening accountability. Scholars of systemic deliberative democracy emphasise that legitimacy arises from the composite health of these channels rather than isolated exchanges. Deepfakes compromise the channels simultaneously, creating a spiral in which each weakened pillar accelerates the decay of the others.

Europe's nascent responses acknowledge this systemic threat but remain partial. Labelling mandates in the AI Act aim to shore up transparency, while proposed platform-researcher partnerships under the European Democracy Action Plan seek to democratise verification capacity, thereby easing equality gaps. Finland's National Media Education Policy (2019) emphasizes systematic media education, quality, and lifelong learning, linking it to societal resilience in the face of disinformation threats (Finland, 2024).

Yet norms must evolve alongside laws. Deliberative legitimacy depends on civic cultures that prize truthful presentation, reciprocal respect, and willingness to be answerable. Technical interventions can scaffold those virtues, but they cannot substitute for them.

Deepfakes expose a vulnerability at the core of democratic architecture, where authenticity functions as a prerequisite for collective self-government. By destabilising what counts as evidence, concentrating communicative power, and enabling strategic denial, synthetic media corrodes the normative pillars that make democracy possible. Regulatory measures may restore partial transparency, and educational programs may narrow literacy gaps, yet democracy ultimately survives on public commitments to truth, equal regard, and responsibility. Reaffirming these commitments in an era of perfect forgeries is not peripheral to technology policy; it is central to democratic renewal.

6 Policy and Educational Responses

Generative-AI systems already create text, images, video, and audio that are almost indistinguishable from authentic material, and the European Commission's Generative AI Outlook warns that such synthetic content could erode public trust during elections and crises if safeguards, including both provenance tracking to verify origin and forensic detection to identify manipulation, do not keep pace (Navajas Cawood et al., 2025). Legislators and regulators are therefore moving from aspirational principles to binding rules that criminalise harmful deepfakes, require visible labelling or watermarking, guarantee rapid takedown mechanisms, limit synthetic political advertising, place detection duties on intermediaries, and oblige model developers to publish transparency reports on training data and risk controls.

Inside the European Union, Article 35 of the Digital Services Act obliges very large online platforms to assess and mitigate systemic risks from manipulated media, label AI-generated content, and give independent researchers secure audit access, with penalties of up to six percent of worldwide turnover for non-compliance. A strengthened Code of Conduct on Disinformation, now formally linked to the Act, extends similar transparency and risk-mitigation duties to search engines and social networks of all sizes and tightens rules on political advertising that uses generative. Forthcoming obligations in the AI Act will reinforce that framework by requiring anyone who publishes synthetic images, audio, or video depicting real people to add notices readable by humans and machines.

Several member states have already gone further. Spain empowered its AI authority to levy fines of up to €35 million, or seven percent of global turnover, on platforms that fail to label synthetic content clearly.[1] France amended its Penal Code to prohibit distributing deepfakes that use a person's likeness or voice without consent unless the artificial origin is disclosed, imposing tougher penalties for sexual material or large-scale online dissemination (Coslin et al., 2024). Commentators note that the new article gives prosecutors a versatile weapon against disinformation campaigns and celebrity impersonations. Germany's Bundesrat circulated a draft Digital

[1] See https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation

Forgery Act that would criminalise synthetic impersonation and introduce higher penalties when victims suffer reputational or economic harm (*Germany: Bundesrat Publishes Draft Law on Deepfakes | News*, 2024). Denmark proposes a copyright-style right in personal biometric features, so reproducing a face or voice in artificial media would require permission or risk infringement liability (Bryant, 2025).

The Italian Constitution safeguards personality rights, including the right to control one's image. Additionally, the Italian Civil Code in its Article 10.5 prohibits the unauthorized use of an individual's likeness, and personal data legislation also protects this. These laws, along with the Italian Copyright Law, enable individuals to seek compensation if their image is used without their consent, especially if it harms their honour or reputation. This clause can arguably be extended to the use of deepfakes. A notable case that exemplifies the enforcement of these laws involved Prime Minister Giorgia Meloni in a lawsuit over pornographic synthetic videos viewed millions of times (Gozzi, 2024). The United Kingdom's Online Safety Act criminalizes both sharing and creating non-consensual intimate deepfakes, with unlimited fines and possible prison sentences (BCC, 2023).

In the United States, the federal landscape remains fragmented, but Congress has introduced the TAKE IT DOWN Act to criminalise non-consensual intimate deepfakes nationwide and compel platforms to provide expedited removal tools (Sen. Cruz, 2025). States continue to fill gaps. Alabama's Child Protection Act treats AI-generated sexual imagery involving minors as virtually indistinguishable from real abuse material (*Alabama HB168*, 2024). California extended its post-mortem right of publicity so that distributing a digital replica of a deceased person without consent triggers civil liability and statutory fines (Wolff & Safran, 2024). Alabama also adopted a Materially Deceptive Election Media statute that outlaws AI-generated content intended to mislead voters (Guidry & Amin, 2024). Arizona clarified that its intimate-image law covers synthetic as well as genuine photographs (Ventura, 2024). Digital Identity Theft Act obliges platforms to host a simple tool for victims, especially minors, to remove explicit deepfakes and criminalizes their non-consensual creation or distribution (*Senator Wahab's Stop the Online Predators Act and Digital Identity Theft Act Signed into Law*, 2024).

Multilateral coordination began to crystallise with the Hiroshima AI Process, whose guiding principles urge developers to publish capability cards, specify disallowed uses, protect intellectual property, and invest in user education so citizens can recognise synthetic media (Japan Gov, 2024). Further to that, the Bletchley Declaration from November 2023 commits signatories to cooperate internationally on the safe development, deployment, and governance of powerful "frontier" AI systems (Department for Science, Innovation and Technology, 2023). Yet implementation is uneven: a European Digital Media Observatory evaluation for the first half of 2024 found that very large platforms met many Code-of-Practice labelling and removal commitments, but smaller services showed limited engagement and inconsistent reporting, underscoring the need for enforcement and capacity-building (Botan & Meyer, 2025).

Because legislation alone cannot keep pace with rapidly improving models, policymakers emphasise technical safeguards and education. A European Parliament briefing on children and deepfakes calls for age-appropriate curricula that teach pupils, parents, and teachers to evaluate digital sources, recognise emotional manipulation, and use verification tools (Negreiro, 2025). The OECD, in cooperation with the European Commission, is drafting an AI-literacy framework that will guide the next Programme for International Student Assessment cycle and provide lesson plans on generative AI (Schleicher, 2025). At the civic level, the EU-funded EUvsDisinfo platform offers an open database of disinformation narratives, interactive games, and instructional videos that help users practise source checking and critical reading (*About - EUvsDisinfo*, 2025).

Research agencies and private companies invest heavily in detection. In the United States, DARPA funds the Semantic Forensics and Media Forensics programmes, which develop algorithms to spot compression artefacts, lighting inconsistencies, and biometric mismatches that indicate tampering (*SemaFor: Semantic Forensics | DARPA*, 2025). Midjourney, a major generative-image service, voluntarily blocks prompts that attempt to create pictures of prominent political figures during election periods, reducing the risk of deceptive visuals entering public debate (O"Brien, 2024). The Massachusetts Institute of Technology's Detect DeepFakes project provides an online training tool where users test their ability to identify manipulated material, and researchers measure how such exercises improve resistance to misinformation.[2] Finland complements these efforts by integrating media-literacy instruction from primary school onward and pairing classroom exercises with

[2] See link: https://detectfakes.media.mit.edu/

public-service broadcasts that explain how manipulated content spreads and how to debunk it (Finland, 2024).

Together, these initiatives raise the cost of deception while preserving the legitimate benefits of generative AI. The European Union's layered strategy, combining horizontal rules like the Digital Services Act with national adaptations and ongoing sector-specific reforms, illustrates how a comprehensive framework can emerge without stifling innovation. In the United States, federal and state measures show that even a patchwork can converge on core principles of consent, transparency and rapid redress. Multilateral dialogues, voluntary industry standards, open-source detection tools and grassroots media-literacy campaigns complete this defence, giving citizens the knowledge and technical support they need to judge what they see and hear before sharing it.

Despite the differences in national approaches, all reviewed examples share the common goal of curbing deepfake abuses and safeguarding the dignity and personal data of citizens. The successful implementation of the European framework (mainly EU AI Act) as a first comprehensive attempt, followed by national adaptations, is expected to lead to more strict enforcement and oversight, and on the flexibility to respond to rapid technological developments. In this context, coordinated international dialogue and the exchange of best practices among Member States are crucial to achieving a balanced and effective regulatory approach that combines innovation with the protection of fundamental human rights.

Beyond formal legislation, industry-led guidelines, technological safeguards, and public awareness campaigns play a vital role in mitigating deepfake risks and promoting responsible AI use. These initiatives, ranging from open-source detection tools to media literacy programs, complement regulatory frameworks by fostering grassroots resilience and rapid adaptation to emerging threats.

Ultimately, a holistic strategy that combines binding rules with voluntary standards and civil-society engagement offers the best path toward an ecosystem where innovation thrives under robust ethical guardrails.

7 Concluding Remarks

Deepfakes pose a structural threat to democratic life by destabilizing the evidentiary foundations of public reason, accountability, and trust. They accelerate the spread of falsehoods while enabling the denial of authentic evidence, creating an epistemic environment in which citizens struggle to distinguish truth from fabrication. These dynamics disproportionately affect women, gender minorities, and other marginalized groups, amplifying social inequalities and deterring full participation in civic and political life. By concentrating communicative power in the hands of technologically sophisticated actors, deepfakes exacerbate inequalities in credibility and reinforce structural hierarchies, undermining the democratic ideal of equal participation.

Addressing these challenges requires a multi-layered approach combining legislation, platform regulation, technological safeguards, and education. Policies such as mandatory labelling, rapid takedowns, and penalties for harmful content, alongside media literacy programs and detection tools, help citizens navigate an increasingly complex information ecosystem. We acknowledge, however, that this analysis is limited by the nascent stage of these regulatory frameworks, whose long-term efficacy in curbing algorithmic disinformation remains to be empirically tested. Yet legal and technical measures alone are insufficient: the resilience of democracy ultimately depends on nurturing civic norms of truthfulness, accountability, and inclusive participation. To this end, future research should prioritize empirical studies that measure the long-term impact of specific media literacy interventions on citizen resilience across diverse political environments.

End notes

Gjon Rakipi conceptualized the chapter, contributed to the abstract, and wrote the sections "Conceptualising Harm" and "Normative Implications for Democratic Values". Additionally, he carried out the preliminary full-chapter edit and assisted in referencing. Andrew McIntyre is the author of "Electoral interference in Europe and Beyond". Calogero Caltagirone and Angelo Tumminelli co-authored "Epistemic Erosion and the Misinformation Ecosystem." Yasaman Yousefi refined the abstract and wrote "Gendered and Minority Harms," drafted the conclusion, performed the final edit, and assisted with the referencing. Asenia Dimitrova authored "Policy and Educational Responses." All authors were individually responsible for the literature review and writing of their respective sections.

References

EUvsDisinfo. (2025, October 27). *EUvsDisinfo.* https://euvsdisinfo.eu/about/

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

Alabama HB168. (2024, May 2). *TrackBill.* https://trackbill.com/bill/alabama-house-bill-168-crimes-offenses-raises-max-age-for-offenses-involving-obscene-materials-with-depictions-of-children-authorizes-punitive-damages-for-victims-of-those-offenses-and-directs-board-of-ed-to-require-policies-related-to-those-offenses/2517763/

Ascott, T. (2020). Microfake: How small-scale deepfakes can undermine society. *Journal of Digital Media and Policy, 11*(2), 215–222.

BBC. (2023, December 9). *AI: EU agrees landmark deal on regulation of artificial intelligence.*https://www.bbc.com/news/world-europe-67668469

Botan, M., & Meyer, T. (2025). *Implementing the EU Code of Practice on Disinformation: An evaluation of VLOPSE compliance and effectiveness (Jan–Jun 2024).* EDMO. https://edmo.eu/publications/implementing-the-eu-code-of-practice-on-disinformation-an-evaluation-of-vlopse-compliance-and-effectiveness-jan-jun-2024/

Bryant, M. (2025, June 27). Denmark to tackle deepfakes by giving people copyright to their own features. *The Guardian.* https://www.theguardian.com/technology/2025/jun/27/deepfakes-denmark-copyright-law-artificial-intelligence

Carpenter, P. (2024). *FAIK: A practical guide to living in a world of deepfakes, disinformation, and AI-generated deceptions.* John Wiley & Sons.

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

Chopra, A., Masroor, G., & Blundy, R. (2025, January 6). "Form of violence": Across globe, deepfake porn targets women politicians. *France 24.* https://www.france24.com/en/live-news/20250106-form-of-violence-across-globe-deepfake-porn-targets-women-politicians

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

Cinelli, M., et al. (2020). The COVID-19 social media infodemic. *Scientific Reports, 10*, 16598.

Coslin, C., Gateau, C., & de Kouchkovsky, A. (2024, July 15). France prohibits non-consensual deep fakes. *Hogan Lovells.* https://www.hoganlovells.com/en/publications/france-prohibits-non-consensual-deep-fakes

Department for Science, Innovation and Technology. (2023, November 1). *The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023.* GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Diakopoulos, N., & Johnson, D. (2019). Anticipating and addressing the ethical implications of deepfakes in the context of elections (SSRN Scholarly Paper No. 3474183). *Social Science Research Network.*https://doi.org/10.2139/ssrn.3474183

European Union Artificial Intelligence Act: A guide. (2025, April 7). *Bird & Bird LLP.* https://www.twobirds.com/-/media/new-website-content/pdfs/capabilities/artificial-intelligence/european-union-artificial-intelligence-act-guide.pdf

Finland Ministry of Education and Culture. (2024, March 12). *Media literacy and education in Finland.* Finland Toolbox. https://toolbox.finland.fi/life-society/media-literacy-and-education-in-finland/

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

Germany: Bundesrat publishes draft law on deepfakes. (2024, July 9). *DataGuidance.*https://www.dataguidance.com/news/germany-bundesrat-publishes-draft-law-deepfakes

Giuffrida, A. (2025, August 28). Outrage in Italy over porn site with doctored images of prominent women. *The Guardian.* https://www.theguardian.com/world/2025/aug/28/outrage-in-italy-over-pornographic-website-with-doctored-images-of-prominent-women-giorgia-meloni

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative language models and automated influence operations: Emerging threats and potential mitigations.* arXiv. https://doi.org/10.48550/arXiv.2301.04246

Gozzi, L. (2024, March 20). Giorgia Meloni: Italian PM seeks damages over deepfake porn videos. *BBC News.*https://www.bbc.com/news/world-europe-68615474

Guidry, T., & Amin, T. (2024, May 24). Alabama takes a stand against AI in political campaigns. *The National Law Review.* https://natlawreview.com/article/new-ai-law-alert-alabama-next-state-take-stand-against-ai-generated-deceptive-media

Hameleers, M., van der Meer, T. G. L. A., & Dobber, T. (2024). They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication, 39.* https://doi.org/10.1177/02673231231184703

Hartmann, T. (2024, April 16). Viral deepfake videos of Le Pen family remind that content moderation is still not up to par ahead of EU elections. *Euractiv.* https://www.euractiv.com/news/viral-deepfake-videos-of-le-pen-family-reminder-that-content-moderation-is-still-not-up-to-par-ahead-of-eu-elections/

Japan Government. (2024, February 9). *The Hiroshima AI Process: Leading the global challenge to shape inclusive governance for generative AI.* https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

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/

Labarre, J. (2025). Epistemic vulnerability: Theory and measurement at the system level. *Political Communication, 42*(1), 6–26. https://doi.org/10.1080/10584609.2024.2363545

Laffier, J., & Rehman, A. (2023). Deepfakes and harm to women. *Journal of Digital Life and Learning, 3*(1), 1–21. https://doi.org/10.51357/jdll.v3i1.218

Levy, N. (2021). Epistemic pollution. In N. Levy (Ed.), *Bad beliefs: Why they happen to good people* (Chap. 5). Oxford University Press. https://doi.org/10.1093/oso/9780192895325.003.0005

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

Meaker, M. (2023, October 3). Slovakia's election deepfakes show AI is a danger to democracy. *WIRED.*https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/

Michael, A., & Hocquard, C. (2023). Artificial intelligence, democracy and elections. *Artificial Intelligence.*

Navajas Cawood, E., Abendroth-Dias, K., Arias Cabarcos, P., Kotsev, A., Bacco, M., Bassani, E., Van Bavel, R., … Sellitto, A. (2025). *Generative AI outlook report: Exploring the intersection of technology, society, and policy.*Publications Office of the EU. https://data.europa.eu/doi/10.2760/1109679

Negreiro, M. (2025). *Children and deepfakes.*

O"Brien, M. (2024, March 13). AI image-generator Midjourney blocks images of Biden and Trump as election looms. *PBS News.* https://www.pbs.org/newshour/politics/ai-image-generator-midjourney-blocks-images-of-biden-and-trump-as-election-looms

O"Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you.* Penguin Press.

Patel, A. (2025). *Freedom of expression, artificial intelligence and elections.* UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000393473

Rini, R., & Cohen, L. (2022). Deepfakes, deep harms. *Journal of Ethics and Social Philosophy, 22*(2). https://doi.org/10.26556/jesp.v22i2.1628

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/

SemaFor: Semantic Forensics | DARPA. (2025, October 27). https://www.darpa.mil/research/programs/semantic-forensics

Sen. Cruz, T. (2025, May 19). *S.146 – TAKE IT DOWN Act (2025–2026)* [Legislation]. https://www.congress.gov/bill/119th-congress/senate-bill/146

Senator Wahab's Stop the Online Predators Act and Digital Identity Theft Act signed into law. (2024, September 19). *California Senate District 10.* https://sd10.senate.ca.gov/news/senator-wahabs-stop-online-predators-act-and-digital-identity-theft-act-signed-law

Swenson, A., Merica, D., & Burke, G. (2024, June 17). AI experimentation is high risk, high reward for low-profile political campaigns. *Associated Press.* https://apnews.com/article/election-2024-ai-deepfakes-political-campaigns-056c2200836e755826fbc9698bcfed60

Tiwari, S. (2024, February 8). Deepfakes become political weapon in Pakistan elections. *India Today.*https://www.indiatoday.in/world/story/deepfakes-become-political-weapon-in-pakistan-elections-2499399-2024-02-08

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3613904.3642382

UNESCO. (2023). *Guidelines for the governance of digital platforms: Safeguarding freedom of expression and access to information through a multistakeholder approach.* https://doi.org/10.54675/OEAJ8758

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

Ventura, G. (2024). The current state of deepfake laws in Arizona. *Arizona State Law Journal.*https://arizonastatelawjournal.org/2024/10/01/the-current-state-of-deepfake-laws-in-arizona/

Weikmann, T., & Lecheler, S. (2024). Cutting through the hype: Understanding the implications of deepfakes for the fact-checking actor-network. *Digital Journalism, 12*(10), 1505–1522. https://doi.org/10.1080/21670811.2023.2194665

Whitmarsh, L. (2011). Responsibility for justice. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001

Wolff, N. E., & Safran, E. (2024, October 30). California expands its post-mortem right of publicity law to cover AI digital replicas. *CDAS.* https://cdas.com/california-expands-its-post-mortem-right-of-publicity-law-to-cover-ai-digital-replicas/

World Health Organization. (2020). *Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation.* WHO Policy Brief.

Yazdani, S., Singh, A., Saxena, N., Wang, Z., Palikhe, A., Pan, D., Pal, U., Yang, J., & Zhang, W. (2025). Generative AI in depth: A survey of recent advances, model variants, and real-world applications. *Journal of Big Data, 12*, 230. https://doi.org/10.1186/s40537-025-01247-x

Young, I. M. (2011). *Responsibility for justice.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001

DOI
https://doi.org/10.18690/um.feri.2.2026.6

ISBN
978-961-299-109-8

Synthetic Media for Social Good: Unlocking Positive Potential

Angelo Tumminelli,[1] Lucy Conover,[2]
Calogero Caltagirone,[1] Giuditta Bassano,[1]
Gjon Rakipi,[3] Tommaso Tonello[2]

[1] LUMSA University, Department of Human Sciences, Rome, Italy
a.tumminelli@lumsa.it, c.caltagirone@lumsa.it, g.bassano@lumsa.it
[2] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
l.a.conover@uu.nl, t.tonello@uu.nl
[3] Albania Institute of International Studies (AIIS), Tirana, Albania
gjonrakipi@aiis-albania.org

This chapter examines the ethical, communicative, and societal dimensions of artificial intelligence for social good (AISG) through a series of participatory workshops conducted in collaboration with the European Citizen Science Association (ECSA). The workshops engaged 44 participants from 18 national backgrounds, selected according to age, gender balance, and domain expertise, and addressed emotionally and epistemically sensitive domains, including climate change communication, the visibility of women in science, and AI-mediated psychological support. The analysis identifies four determinants shaping perceived impact: narrative–intentional coherence, technical–mimetic realism, ethical transparency, and contextual adequacy. Together, these dimensions inform a preliminary set of ethical and design guidelines for socially engaged and educational media. The chapter further proposes a methodological framework that combines semiotic modelling with iterative user testing to evaluate AI-generated content beyond criteria of realism or imitation. By foregrounding communicative function, ethical clarity, and cultural resonance, the findings suggest that synthetic media can meaningfully contribute to socially oriented and educational contexts when designed with participatory and ethically grounded approaches.

Keywords:
social good,
literacy,
citizen science,
education,
ethics

University of Maribor Press

1 Why AI for Good?

The purpose of ethics is to promote the full flourishing of people in their deepest relational openness and in their aspiration to meaning. Ethics of AI is, therefore, called to not only define the normative criteria within which to place the interaction between AI and human beings, but above all, to identify the strategies with which the use of the former is placed at the service of personal fulfilment and the common good. Thus, the ethics of AI goes well beyond a merely deontological approach, constituting itself, rather, as a fundamental tool for promoting human beings in the face of the challenges imposed by the digital revolution and the advent of AI.

We follow Aristotle, who in the *Nicomachean Ethics* argued that within society, the common good must be pursued as a supreme ethical task to which individual action is called to contribute significantly. According to Aristotelian teleology, every being is oriented toward an end (*telos*) and evaluates actions based on how well they realize the human good. (Aristotle, 2012, I, 1094a, pp. 1–3)

In contemporary AI ethics, this idea reappears when defining the desirable goals of intelligent systems and the criteria for judging their alignment with human values. In both perspectives, what matters is determining which end should guide action to direct AI development and use toward the common good.

In the present-day debate, Luciano Floridi also explains the potential of the political use of AI for the common, or social good (AI4SG), highlighting how its ethical use necessarily implies "the design, development and implementation of AI systems in order to (I) prevent, mitigate or solve problems that negatively impact human life and/or the well-being of the natural world and/or (II) allow socially preferable and/or environmentally sustainable developments" (Floridi, 2022, p. 223).

Luciano Floridi (with Josh Cowls) proposes five fundamental ethical principles for AI, often referred to as the "Unified Framework of AI Ethics":

Beneficence – AI should promote well-being and generate social value.

Non-maleficence – it should avoid harm, undue risks, and abusive uses.

Autonomy – it should respect individuals' decision-making capacity without manipulating them.

Justice – it should be fair, non-discriminatory, and distribute benefits and burdens appropriately.

Explicability – it should ensure transparency, intelligibility, and traceability of decisions.

Floridi bases his reflection on the ethics of AI on these five principles, borrowed from an accredited approach in bioethics, to combine the use of AI and the promotion of the individual and the common good of humanity. In compliance with the principle of Beneficence, according to Floridi, it is necessary to create an AI technology that is beneficial for humanity and that puts the promotion of the well-being of people and the planet at its centre, thus safeguarding the human dignity of the present and the future as a common good.

The principle of non-maleficence, on the other hand, is based on the need to prevent violations of personal privacy to avoid improper use of AI technologies that could harm humanity as a whole. The principle of Autonomy, then, is the one that is called to safeguard the freedom of individuals as a shared heritage (Floridi, 2022): if it is true that when AI and its intelligent action are adopted, the individual voluntarily gives up part of his decision-making power to machines, affirming the principle of Autonomy in the context of AI means reaching a balance between the decision-making power that the individual retains within himself and that which he delegates to artificial agents. Starting from this, not only should human freedom be promoted, but also the autonomy of machines should be restricted and made intrinsically reversible.

Floridi's perspective is particularly interesting because it places the social good and the possibility that it can be achieved through personal freedoms at the centre of an ethical use of AI (Floridi, 2022; Floridi et al., 2020). Only when this happens in a society can the common good be achieved: this is not a utopia but an ethical task that awaits all human beings in the face of the challenges of their time.

If ethics aims to guide human action toward personal flourishing and meaningful relationships, then AI ethics must not only set the norms governing human-AI interaction, but also determine how AI can genuinely support human fulfilment and the common good. Thus, AI ethics goes beyond a purely deontological framework: it becomes a key instrument for fostering human development in the face of the digital revolution and the rise of AI.

2 Positive Applications in Citizen Science, Community Engagement, and Education

Since 2015, the United Nations Sustainable Development Goals (UNSDGs) have been endorsed by all UN Member States to tackle the most pressing social, environmental, and economic issues by 2030. Citizen science, as "a form of research collaboration involving members of the public in scientific research projects to address real-world problems" (Wiggins & Crowston, 2012) has proven its contribution to the SDGs. Citizen science is an "umbrella term" to include various participatory approaches where non-professional scientists contribute to research (ECSA, 2015; 2020), such as participatory monitoring, crowd-sourced science, or participatory action research. Indeed, participatory approaches leveraging public involvement have demonstrated to significantly enhance data collection, foster community empowerment, and drive progress toward achieving the SDGs (Ballerini & Bergh, 2021; Fraisl et al., 2023; Gaventa & Barrett, 2012; Huttunen et al., 2022; Loeffler & Martin, 2015; Müller et al., 2023). In this section, we show how AI is used in citizen science initiatives, community engagement and education to support the Sustainable Development Goals. This section will present a short background of different types of AI-supported citizen science initiatives and learnings from the SOLARIS project, which constitute the bedrock of the activities carried out during Use Case 3 (UC3).

In citizen science, AI-driven tools can enhance data analysis, pattern recognition, and predictive modelling, not only improving the efficiency and accuracy of citizen science projects, but also expanding their scope and scalability (Fraisl et al., 2025; Hayes et al., 2025; Sinha et al., 2024). Among citizen science projects, the most common way of integrating AI is by having participants train algorithms (Chandler et al., 2025; DeSpain et al., 2024; Duerinckx et al., 2024, p. 3; Jia et al., 2025; See et al., 2025). This is sometimes called "hybrid intelligence (HI) systems" (Chen et al., 2024) or "Crowd AI" (Palmer et al., 2021), as citizen scientists provide data and support machine classification tasks, for example in monitoring efforts such as high-tide flooding (Golparvar & Wang, 2020), vector-borne diseases (Saran & Singh, 2024), or harmful mosquitos or snails (Chan et al., 2024). AI use in citizen science also enhances challenges such as the mitigation of algorithmic biases (Vinuesa et al., 2020) and inclusive, accessible technological designs that ensure broad participation (Fortson et al., 2024). Questions remain in terms of data privacy, hence emphasizing the importance of adopting ethical frameworks that prioritize transparency, accountability, and fairness in citizen science projects (Ceccaroni et al., 2019; Fortson et al., 2024; Vinuesa et al., 2020). In citizen science biodiversity research, for instance, AI can be used for species identification (Hogeweg et al., 2024), such as mammal species in the FOOTPRINTS-CITSC project,[1] or diseases on potato crops in the PataFest project.[2] Additionally, AI chatbots on biodiversity monitoring platforms have also been shown to enhance engagement, as contributors use the bot as a "dialogic partner" to discuss the pictures of bumblebees they upload (Sharma et al., 2024). And yet, power asymmetries in current data governance still fail to properly acknowledge citizen scientists as relevant stakeholders for drafting and implementing data principles, which in turn inform data storage and data use. Nonetheless, the same public engagement values that support citizen science would appear to benefit ethical data governance: there already exist positive initiatives, especially in relation to citizen science as undertaken within indigenous communities, to inquire into local knowledge. By fostering data justice processes – e.g., through the promotion of data commons and cooperatives – and the enhancement of multi-stakeholder data governance processes through its participatory principles, citizen science represents a relevant tool to also enhance accountability mechanisms and to democratise data governance (Borda & Greshake Tzovaras, 2025; Sterner & Elliott, 2024). In the educational sector, the Smartschool project,[3] Supporting teachers and pupils through a smart signal, is currently working on an AI tool for teachers to identify their teenage students' learning needs on a learning platform. The project is a collaboration between students, parents, education professionals, and Hasselt University.[4] Furthermore, the Monumai project[5] citizens participate in data collection and training algorithms to recognize architectural styles from photographs of monuments, whereby they also learning to recognize the characteristics. In the care sector, the project "Machine learning as a citizen science tool to improve the quality of life of older people and their caregivers'[6] wants to make psychology and computer science research accessible to

[1] See link: https://footprints.citizenscience.no/

[2] See link: https://www.patafest.eu/

[3] See link: https://citizenscience.eu/project/488

[4] See link: https://www.uhasselt.be/en/faculties-and-schools/school-of-social-sciences

[5] See link: https://monumai.ugr.es/

[6] See link: https://citizenscience.eu/project/72

the wider society and support the early detection of loneliness, social isolation, and stress in older adults. Data is provided by volunteers, who will analyse it before feeding machine learning algorithms for training.

The aforementioned projects show how, across disciplines, citizen science initiatives are increasingly using AI tools to address various SDGs. "AI for good", in the context of UC3, means AI to achieve the SDGs. By promoting citizens' participation in the co-creation of AI-generated content for educational purposes, UC3 aimed to promote AI to achieve the SDGs, or "AI for good". It supported SDG 4 - Quality Education, in two ways: first, participants co-created content for awareness raising – on topics such as climate change; second, the workshops fostered participants' digital literacy and enabled individuals to better understand and navigate the complexities of AI technology. UC3 also played a significant role in advancing SDG 16 - Peace, Justice, and Strong Institutions, by pushing for pro-democratic values and promoting transparency and accountability in AI governance. The participatory governance model inherent in UC3 encouraged citizens to take an active role in decision-making processes, thereby ensuring that AI systems align with societal values. In practice, we selected three SDGs to promote "AI for good":

- SDG 3: Good Health and well-being, focusing on mental health,
- SDG 5: Gender equality, especially with regards to the inclusion of women in science, and
- SDG 13 Climate Action, focusing on the effects of climate change.

SOLARIS project member created eight videos on these themes. During the workshops part of SOLARIS UC3 activities, we therefore contributed to an acceptable or desirable approach for awareness raising of artificially generated content. We framed possible answers to the question: "what could "good" AI-generated content look like?" By enabling citizens to co-create AI-generated content with experts, the workshops contributed to the transparency, inclusivity, and accountability that are fundamental to democratic governance. The workshops were also based on the value-sensitive design approach (Umbrello & Van De Poel, 2021, p. 284), which takes "values of ethical importance into account", considering "a tripartite methodology of empirical, conceptual and technical investigations".

3 Semiotic at the service of AI for Good

Use Case 3 explored the civic and communicative potential of “positive deepfakes,” that is, synthetic texts generated by AI for educational, memorial, scientific, and civic engagement purposes, rather than for manipulative or deceptive purposes. UC3 adopted a semiotic and processual approach. Its goal was not to evaluate persuasion or misinformation, but to understand how artificial texts[7] are constructed, which dimensions guarantee their credibility, or conversely, reveal their artificiality, and how workshop participants interpret such products by attributing meaning to them.

Within this framework, “semiotics”, understood as the science of meaning-making forms and of the conditions of their production and interpretation (Eco, 1976; Greimas, 1983; Greimas & Courtés, 1982; Hjelmslev, 1961) was considered a useful framework to complement the ethical perspective of AI4SG. UC3, therefore, sought to approach deepfakes as semiotic objects whose analysis requires decomposition into levels of textual articulation and reconstruction of the pragmatic conditions of reception. Hence, there is a need for a multilevel analysis integrating discursive, narrative, enunciative, axiological, and plastic components to map how synthetic contents acquire meaning and produce social effects. From a semiotic perspective, each artificially generated video can be analysed as a text articulated on multiple levels:

- Discursive level: any audiovisual text, even a static one, “speaks” of something, projects figures, situates them in space and time, and constructs a coherent discursive universe.
- Narrative level: concerns the characters’ actions, the transformations that occur, and the evolution of the storyline. It is the level at which conflicts, changes of state, and narrative programs can be observed.
- Enunciative level: includes the traces indicating the relationship between sender and receiver, the contracts of truth, and the framing regimes (fiction, testimony, document, hybrid, etc.).
- Axiological level: relates to the explicit or implicit values conveyed by the text, such as truth, authority, empathy, transparency, or responsibility.

[7] In semiotics, “text” is a generic term that can refer to audiovisual contents too.

To these levels, we add the specificity of visual and audiovisual texts. According to Polidoro (2008), visual semiotics distinguishes two areas of analysis:

- Figurative semiotics, which analyses meaning derived from the recognition of objects and scenes.
- Plastic semiotics, which investigates the significance of visual configurations such as shapes, colours, textures, and lighting.

This dual articulation suggests that the plausibility of visual content does not depend solely on perceptual accuracy but is mediated by cultural codes and cognitive competencies. Visual literacy is built over time through familiarity with communicative genres, aesthetic codes, and narrative conventions. This was the ground on which UC3 developed its investigation.

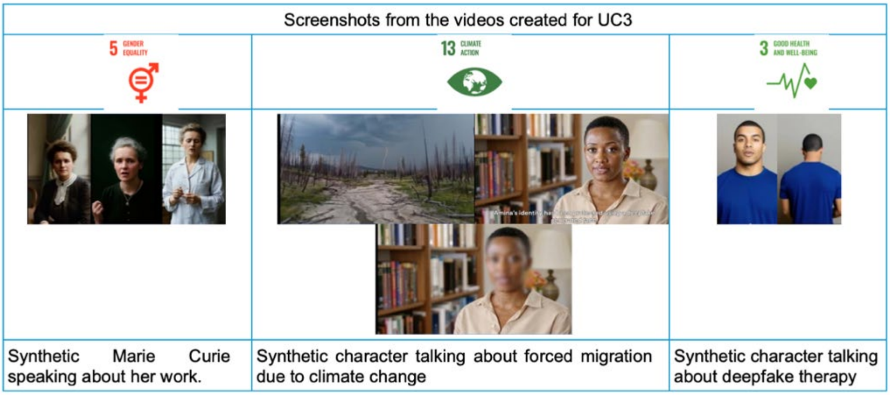
The eight videos produced in UC3 were designed to systematically and creatively test a set of variables.[8] The language used in all videos was English, and the videos covered the following themes:

- **SDG3: Women Scientists– Marie Curie**: three videos presented the scientist as an authoritative witness, capable of reflecting on the role of women in science.
- **SDG5: Climate Crisis– Amina:** Two videos narrated the experience of a woman forced to leave her homeland near Lake Chad due to desertification.
- **SDG13: Mental Health– Casey**: Two videos explored the use of synthetic avatars in psychological therapy.

The design logic was to combine predefined variables to observe thresholds of acceptability and mechanisms of suspended disbelief. Eight variables were initially identified, derived from narratological frameworks already adapted in previous research on synthetic media and video analysis (Bassano & Cerutti, 2024; Genette, 1982; Greimas, 1988). Their articulation allowed us to operationalise classical narrative dimensions, namely actoriality, focalization, setting, and modality within an experimental design suited to AI-generated content:

[8] The videos are safely stored in the SOLARIS archive and can be accessed upon request, but they are not publicly accessible.

1) famous vs. an anonymous person.
2) realistic vs. decontextualized/abstract setting.
3) monologue vs. dialogue.
4) focus on detail vs. overall view.
5) blurred face vs. AI-generated (deepfake) face.
6) first-person vs. third-person narration.
7) artificial landscape vs. artificial person.
8) serious vs. entertainment context.


Screenshots from the videos created for UC3
5 GENDER EQUALITY
13 CLIMATE ACTION
3 GOOD HEALTH AND WELL-BEING

Synthetic Marie Curie speaking about her work.	Synthetic character talking about forced migration due to climate change	Synthetic character talking about deepfake therapy

Figure 6.1: Screenshots from the videos created for UC3
Source: SOLARIS

For practical reasons, the deepfakes focused on five of these variables (1, 2, 5, 6, 7), which were articulated across the three themes described above. The scripts were initially proposed by ECSA, then further developed and conceptually authored by Giuditta Bassano (LUMSA), and finally produced by the partner CINI, in particular by Michele Brienza.

3.1 The Textual Taxonomy of UC3

Based on this theoretical framework, and on the analysis of data collected during the workshops, we propose a textual classification of the positive deepfakes used during UC3 along three principal axes: (i) their discursive form, (ii) their identity function, and (iii) their destination. These three axes, intertwined with one another, enable the distinction of how synthetic actors acquire meaning and produce communicative effects. This taxonomy, specifically developed for the purposes of this project and constituting an original contribution of this chapter, indicates that the evaluation of *positive* AI-generated contents cannot be based solely on technical quality. Instead, they must be read as complex textual configurations capable of combining different degrees of discursive involvement, identity strategies, and forms of destination. In this section, the term "textual" refers to the intrinsic configuration of the deepfake as a discursive object: its narrative structure, identity work, and intended destination. This level concerns the organization of meaning within the text itself, independently of how it is received. By contrast, the interpretive taxonomy presented in the following section focuses on the modes of reception activated by audiences, showing how viewers make sense of the same textual features through different perceptual, cognitive, and ethical frameworks.

The first axis (discursive form) concerns the degree of personal involvement that the narrator assumes in the account. We can imagine a continuous spectrum with two opposite poles. On one side, we would place the evocative or illustrative pole. This occurs when the narrative voice remains external, minimally engaged in the first person, limiting itself to evoking facts or presenting issues. This is the case of Marie Curie: even when referring to her own biography, the scientist appears rather detached, informing us of "public" events, already known and of common interest, thus functioning more as an exemplary figure than as a subject testifying in the first person to a personal experience. On the opposite side, we find the testimonial pole, a position that entails the highest degree of intimacy and subjective implication. Casey's narrative could have been placed here, especially if the synthetic actor had gone so far as to describe concrete details of his anxiety disorder.

The second axis (identity function) concerns the way in which deepfakes handle the identity of the subject being represented. We distinguish between passive and active functions. The passive function consists in covering and protecting a real identity by concealing its individual traits. This is the case of Amina and Casey, whose faces were blurred or withheld from view, to safeguard anonymity or reduce exposure. The active function, instead, corresponds to the maximum degree of identity affirmation, when the deepfake serves a memorial function, bringing historical figures back to life to prolong their presence. This is the case of Marie Curie, who appears or is evoked in the three videos as a historical and symbolic figure, whose identity is not concealed but reaffirmed and consolidated.

The third axis (destination) concerns intended use of deepfakes. Here, too, we can imagine a continuum. On one end lies the public pole, meaning texts designed for a broad, general audience, such as the Dalí deepfake (evoked during the UC3 workshops) in a museum setting. The videos of Marie Curie also share this orientation: they are meant to convey collective values and educational messages. On the other end lies the specific pole, which refers to contents designed for situated, personalized, or dialogic use. This is the case of the videos about Casey, which evoke an individual therapeutic context, as well as the workshop discussions about chatbots as personal assistants capable of establishing a unique relationship with a single user. By combining the three axes, it is possible to position the UC3 cases within a textual matrix:

- Marie Curie: *evocative, active, public*;
- Amina: *evocative/testimonial, passive, public*;
- Casey: *testimonial, passive, specific*;

Considered together, the three cases display different types of balance across the proposed axes. Marie Curie, as a historical and already public figure, clearly occupies an evocative position on the first axis, rather than a testimonial one, since the narrative mobilizes shared and well-known events without direct personal involvement. On the second axis, her deepfake performs an active identity function, reinforcing and extending her symbolic presence. Finally, its destination is unmistakably public, oriented toward broad educational dissemination. Amina occupies a more nuanced position: her discourse is predominantly evocative, yet certain passages introduce elements of testimonial engagement. Her identity, however, remains passively configured, as the message protects and obscures individual traits; her destination is likewise public, given that the content is framed as a general appeal. Casey stands at the opposite corner of the matrix: his deepfake is grounded in a strongly testimonial mode, openly engaging personal experience; his identity is passive, since his face is concealed for privacy reasons; and the destination is specific, as the video aligns with therapeutic or relational contexts rather than with broad public dissemination.

3.2 The Interpretive Taxonomy

While the textual taxonomy has made it possible to classify civic deepfakes according to their formal and discursive configuration, an interpretive taxonomy allows us to understand their modes of reception. The UC3 workshops showed that the credibility of deepfakes does not depend solely on technical realism but unfolds through different interpretive registers activated by the audience when encountering the texts. We can distinguish five primary modes of reception:

1) *Plastic interpretation*: this is the most immediate threshold of access, linked to visual and auditory perception. Details such as lip-sync, frame rate, coherence of lighting and textures, movement rhythm, or the quality of the synthetic voice constitute decisive clues for acceptance or rejection. In the workshops, younger participants proved particularly sensitive to this level: for them, plastic realism represented a non-negotiable condition of credibility. This emerged clearly in reactions to Marie Curie's slightly imperfect lip-sync, which younger participants immediately flagged as a credibility break.

2) *Discursive interpretation*: beyond the plastic level, viewers assessed the content based on narrative and thematic coherence. Here, the effects of meaning emerge, tied to the construction of plausible stories, the consistency of the conveyed values, and the text's ability to articulate a meaningful account. Older participants tended to prioritize this dimension, paying greater attention to the quality of discourse than to technical perfection. For instance, when the video on climate-change consequences was shown, participants focused on the coherence between the verbal text and the visual depiction of environmental impacts.

3) *Ethical-cognitive interpretation*: the reception of civic deepfakes also implies a judgment about the appropriateness of their use in specific contexts. The workshops revealed that a deepfake may be deemed acceptable in a museum or classroom, yet disturbing in a promotional or commercial setting. This level thus concerns the audience's ability to relate synthetic content to social and ethical frameworks, evaluating its legitimacy and transparency. For example, in Casey's case, participants noted that it would be inappropriate to use an avatar of someone with mental health disorders in a pharmaceutical advertisement or in promotional material for medical services. They also stressed, however, that

this is very different from the experience of a patient with mental health conditions who wants to educate and inform others through a deepfake.

4) *Passional interpretation*: a fourth register concerns the emotional dimension. Reception depends on the alignment between sensible form and narrated content: a smiling face recounting a trauma generates discomfort, whereas an empathetic tone strengthens the text's acceptability. This aspect became evident when participants discussed the quality of Amina's video, noting that her expression appeared too cheerful compared to the dramatic nature of what she was describing.

5) *Metareflective interpretation*: finally, a more sophisticated mode arises when participants thematize the deepfake itself as an object of reflection. Co-creation fostered this level: citizens discussed the contents and the cultural, ethical, and political implications of the technology, highlighting their active role as critical interpreters. This mode emerged directly from the workshop discussions, as a recurrent interpretive pattern observed among participants. In UC3, this mode surfaced when participants discussed the broader implications of using deepfakes of figures like Marie Curie, Amina, and Casey in civic contexts.

The intersection between the textual and interpretive taxonomies shows how the three strands of UC3 were received in different ways. For Marie Curie, the public dimension seemed to strengthen acceptability, even though workshop participants still emphasized discursive and ethical-cognitive interpretation (given the educational context). For the synthetic character of Amina, identity protection and blurring weakened the testimonial effect; participants oscillated between plastic rejection (the synchronization of body and facial movements was judged unconvincing) and passional discomfort, while nevertheless paying attention to significant metareflective aspects, such as the synthetic actress's voice. For the synthetic character of Casey, the testimonial effect appears to have failed altogether, as participants mainly interpreted the video in plastic and passional terms, discussing evident artificiality and a sense of detachment. The analysis of the workshops provided a rich picture of how citizens interpret and evaluate synthetic content, offering empirical validation for the two taxonomies developed. The results extend beyond observing individual reactions, as they demonstrate how participants employed complex interpretive strategies, combining plastic, discursive, ethical-cognitive, passional, and metareflective evaluations.

Despite the richness of its findings, UC3 presents certain structural limitations tied to the online workshops' format. The videos were shown in standardized, decontextualised conditions, far removed from the communicative ecosystems in which synthetic content circulates typically. As already noted, a deepfake never exists in isolation: its meaning depends on the discourses that accompany it, the users' comments, the platforms that host it, the viewing devices, and the intertextual frameworks into which it is inserted – this is the network approach developed by SOLARIS project (see McIntyre et al., 2025, Bisconti et al., 2024).

4 Concluding Remarks

Our findings bring to the fore the theme of "Digital education". Digital education plays a crucial role in developing skills for digital citizenship and democracy, as it trains individuals capable of interacting consciously, responsibly, and actively in a digital context. These skills are essential to navigate the online world and to participate in democratic life with critical thinking and respect, promoting open and inclusive dialogue. Digital education promotes skills such as critical thinking, responsibility, respect for privacy and digital rights, the fight against disinformation, and active participation. In this regard, starting from the interplay between empirical findings and theoretical models, Panciroli and Rivoltella (2023) speak of "algorithmic pedagogy", meaning the set of strategies that make use of technological and digital devices used in educational contexts to promote learning and the integral formation of the person. The two scholars refer to three possible configurations of algorithmic pedagogy, and distinguish: 1. "AI in education", which involves the teacher being supported by a humanoid robot available to answer students' questions based on profiling and individualized programming processes (here the reference is to robots used in co-teaching for feedback management and personalized tutoring); 2. "AI by education", or the provision of pre-established and predetermined ethical criteria for devices in the design phase (in this regard, the responsibility of the computer designer comes into play, who, already in the creation of the algorithm and in the writing of the code, establishes limits and ethical criteria); 3. "AI for education", which consists of the task of digital education, aimed at arousing critical thinking in students. This awareness implies distancing from the technological artefact, which is recognized in its functional utility and not as a substitute for interpersonal educational relationships. An ethical digital education, in the context of the infosphere, thus becomes an essential basis for the promotion of humanity and the construction of the common good.

Overall, we see that AI has the potential to promote social good, if it is developed and used responsibly. By maintaining thoughtful reflection about the complexities of AI in the context of education and social good, the technology could be used to provide a positive lens in these fields. However, future endeavours need to avoid the deficit model, which considers the general public as only lacking skills to interact with AI: while education has a crucial role to play, focusing only on digital education tends to reinforce systemic barriers to participation and inclusion (Patel, 2025). Instead, we need to ensure that diverse voices are included and can participate in the development of tools and technologies influencing society. Future research should focus on participatory co-design of educational AI tools.

End notes

Angelo Tumminelli and Calogero Caltagirone conceptualized the chapter, coordinated the writing of the text, and conducted the literature review. Angelo Tumminelli and Calogero Caltagirone wrote the abstract, conclusions, and section "Why AI for good?". Gjon Rakipi and Tommaso Tonello wrote Section "Positive applications in citizen science, community engagement and education"; Giuditta Bassano wrote Sections "semiotic at the service of AI for Good", "The Textual Taxonomy of UC3", and "The Interpretive Taxonomy". Lucy Conover harmonized the whole text and coordinated revisions, including reference formatting and editing for grammar and style of the entire chapter. All authors reviewed and approved the final version. We thank the colleagues from the European Citizen Science Association (ECSA) – Romane Blassel, Nefertari Nachtigall, and Carolina Doran for helpful comments and suggestions on earlier versions of this chapter, and for their contributions to the activities described and discussed in this chapter.

References

Aristotle. (2012). *Nicomachean ethics* (R. C. Bartlett & S. D. Collins, Trans.). University of Chicago Press.

Ballerini, L., & Bergh, S. I. (2021). Using citizen science data to monitor the Sustainable Development Goals: A bottom-up analysis. *Sustainability Science, 16*(6), 1945–1962. https://doi.org/10.1007/s11625-021-01001-1

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. https://doi.org/10.1007/s11196-023-10067-2

Bisconti, P., McIntyre, A., & Russo, F. (2024). Synthetic socio-technical systems: Poiêsis as meaning making. *Philosophy & Technology, 37*(3), Article 94. https://doi.org/10.1007/s13347-024-00778-0

Borda, A., & Greshake Tzovaras, B. (2025). *Perspectives on crowdsourced citizen science and the data governance of AI applications* (SSRN Scholarly Paper No. 5291262). Social Science Research Network. https://doi.org/10.2139/ssrn.5291262

Ceccaroni, L., Bibby, J., Roger, E., Flemons, P., Michael, K., Fagan, L., & Oliver, J. L. (2019). Opportunities and risks for citizen science in the age of artificial intelligence. *Citizen Science: Theory and Practice, 4*(1), Article 29. https://doi.org/10.5334/cstp.241

Chan, K. H., Tumusiime, J., Jacobs, L., & Huyse, T. (2024). The potential of deep learning object detection in citizen-driven snail host monitoring to map putative disease transmission sites. *Citizen Science: Theory and Practice, 9*(1), Article 25. https://doi.org/10.5334/cstp.724

Chandler, C. O., Sedaghat, N., Oldroyd, W. J., Frissell, M. K., Trujillo, C. A., Burris, W. A., Hsieh, H. H., Kueny, J. K., Farrell, K. A., Borisov, G. V., DeSpain, J. A., Bernardinelli, P. H., Kurlander, J., Magbanua, M. J. M., Sheppard, S. S., Mazzucato, M. T., Bosch, M. K. D., Shaw-Diaz, T., Gonano, V., … Dukes, C. J. A. (2025). AI-enhanced citizen science discovers cometary activity on near-Earth object (523822) 2012 DG61. *Research Notes of the AAS, 9*(1), Article 3. https://doi.org/10.3847/2515-5172/ada368

Chen, V. Y., Lu, D.-J., & Han, Y.-S. (2024). Hybrid intelligence for marine biodiversity: Integrating citizen science with AI for enhanced intertidal conservation efforts at Cape Santiago, Taiwan. *Sustainability, 16*(1), 454. https://doi.org/10.3390/su16010454

DeSpain, J. A., Chandler, C. O., Sedaghat, N., Oldroyd, W. J., Trujillo, C. A., Burris, W. A., Hsieh, H. H., Kueny, J. K., Farrell, K. A., Magbanua, M. J. M., Sheppard, S. S., Mazzucato, M. T., Bosch, M. K. D., Shaw-Diaz, T., Gonano, V., Lamperti, A., Da Silva Campos, J. A., Goodwin, B. L., Terentev, I. A., & Dukes, C. J. A. (2024). Discovery of Jupiter family comet 2011 UG104 through AI-enhanced citizen science. *Research Notes of the AAS, 8*(5), Article 140. https://doi.org/10.3847/2515-5172/ad4d9c

Duerinckx, A., Veeckman, C., Verstraelen, K., Singh, N., Van Laer, J., Vaes, M., Vandooren, C., & Duysburgh, P. (2024). Co-creating artificial intelligence: Designing and enhancing democratic AI solutions through citizen science. *Citizen Science: Theory and Practice, 9*(1), Article 43. https://doi.org/10.5334/cstp.732

Eco, U. (1976). *A theory of semiotics.* Indiana University Press.

Floridi, L. (2022). *Etica dell"intelligenza artificiale: Sviluppi, opportunità, sfide* (M. Durante, Ed.; 1st ed.). Raffaello Cortina Editore.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics, 26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Fortson, L., Crowston, K., Kloetzer, L., & Ponti, M. (2024). Artificial intelligence and the future of citizen science. *Citizen Science: Theory and Practice, 9*(1), Article 32. https://doi.org/10.5334/cstp.812

Fraisl, D., See, L., Fritz, S., Haklay, M., & McCallum, I. (2025). Leveraging the collaborative power of AI and citizen science for sustainable development. *Nature Sustainability, 8*(2), 125–132. https://doi.org/10.1038/s41893-024-01489-2

Gaventa, J., & Barrett, G. (2012). Mapping the outcomes of citizen engagement. *World Development, 40*(12), 2399–2410. https://doi.org/10.1016/j.worlddev.2012.05.014

Genette, G. (1982). *Figures of literary discourse* (A. Sheridan, Trans.). Columbia University Press.

Golparvar, B., & Wang, R.-Q. (2020). AI-supported citizen science to monitor high-tide flooding in Newport Beach, California. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities* (pp. 66–69). Association for Computing Machinery. https://doi.org/10.1145/3423455.3430315

Greimas, A. J. (1983). *Structural semantics: An attempt at a method* (R. Schleifer, D. McDowell, & A. R. Velie, Trans.). University of Nebraska Press.

Greimas, A. J. (Ed.). (1988). *Maupassant: The semiotics of text: Practical exercises* (P. Perron, Trans.). John Benjamins Publishing Company. https://doi.org/10.1075/sc.1

Greimas, A. J., & Courtés, J. (Eds.). (1982). *Semiotics and language: An analytical dictionary.* Indiana University Press.

Hayes, S., Jandrić, P., la Velle, L., Earle, S., Šrajer, F., Dragić, Z., Kubat, S., Peraica, A., Švraka, D., Popović, S., Mumelaš, D., Pospiš, D., Vujanović, B., Lugović, S., Jopling, M., Tolbert, S., & Watermeyer, R. (2025). Postdigital citizen science and humanities: Dialogue from the ground. *Postdigital Science and Education, 7*(1), 188–223. https://doi.org/10.1007/s42438-024-00514-z

Hjelmslev, L. (1961). *Prolegomena to a theory of language* (F. J. Whitfield, Trans.; Rev. English ed.). University of Wisconsin Press.

Hogeweg, L., Yan, N., Brunink, D., Ezzaki-Chokri, K., Gerritsen, W., Pucci, R., Ghani, B., Stowell, D., & Kalkman, V. (2024). AI species identification using image and sound recognition for citizen science, collection management and biomonitoring: From training pipeline to large-scale models. *Biodiversity Information Science and Standards, 8*, e136839. https://doi.org/10.3897/biss.8.136839

Huttunen, S., Ojanen, M., Ott, A., & Saarikoski, H. (2022). What about citizens? A literature review of citizen engagement in sustainability transitions research. *Energy Research & Social Science, 91*, 102714. https://doi.org/10.1016/j.erss.2022.102714

Jia, P., Lv, J., Li, Y., Song, Y., Fu, M., Li, Z., Liu, Y., Li, N., Shen, S., Wang, T., Li, R., Zhou, Z., Ren, J., He, Z., Li, S., Tao, Y., & Cui, C. (2025). Galaxy Circus: A new paradigm for anomalous galaxy discovery with artificial intelligence and citizen science. *Preprint in review*. https://doi.org/10.21203/rs.3.rs-6397547/v1

Loeffler, E., & Martin, S. (2015). Citizen engagement. In *Public management and governance* (3rd ed.). Routledge.

McIntyre, A., Conover, L., & Russo, F. (2025). A network approach to public trust in generative AI. *Philosophy & Technology, 38*(4), Article 137. https://doi.org/10.1007/s13347-025-00974-6

Müller, M., Lorenz, J., Voigt-Heucke, S., Heinrich, G., & Oesterheld, M. (2023). Citizen science for the Sustainable Development Goals? The perspective of German citizen science practitioners on the relationship between citizen science and the Sustainable Development Goals. *Citizen Science: Theory and Practice, 8*(1), Article 34. https://doi.org/10.5334/cstp.583

Palmer, M. S., Huebner, S. E., Willi, M., Fortson, L., & Packer, C. (2021). Citizen science, computing, and conservation: How can "crowd AI" change the way we tackle large-scale ecological challenges? *Human Computation, 8*(2), 54–75. https://doi.org/10.15346/hc.v8i2.123

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

Polidoro, P. (2008). *Che cos'è la semiotica visiva* (1st ed.). Carocci.

Saran, S., & Singh, P. (2024). Systematic review on citizen science and artificial intelligence for vector-borne diseases. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-4–2024*, 397–402. https://doi.org/10.5194/isprs-archives-XLVIII-4-2024-397-2024

See, L., Chen, Q., Crooks, A, Laso Bayas, J. C., Fraisl, D., Fritz, S., Georgieva, I., Hager, G., Hofer, M., Lesiv, M., Malek, Ž., Milenković, M., Moorthy, I., Orduña-Cabrera, F., Pérez-Guzmán, K., Schepaschenko, D., Shchepashchenko, M., Steinhauser, J., & McCallum, I. (2025). New directions in mapping the Earth's surface with citizen science and generative AI. *iScience, 28*(3), 111919. https://doi.org/10.1016/j.isci.2025.111919

Sharma, N., Colucci-Gray, L., Lakeman-Fraser, P., Robinson, A., Newman, J., Van der Wal, R., Rueger, S., & Siddharthan, A. (2024). Image recognition as a "dialogic AI partner" within biodiversity citizen science – An empirical investigation. *Citizen Science: Theory and Practice, 9*(1), Article 35. https://doi.org/10.5334/cstp.735

Sinha, R. K., Kumar, R., Phartyal, S. S., & Sharma, P. (2024). Interventions of citizen science for mitigation and management of plastic pollution: Understanding sustainable development goals, policies, and regulations. *Science of The Total Environment, 955*, 176621. https://doi.org/10.1016/j.scitotenv.2024.176621

Sterner, B., & Elliott, S. (2024). How data governance principles influence participation in biodiversity science. *Science as Culture, 33*(3), 366–391. https://doi.org/10.1080/09505431.2023.2214155

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics, 1*(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3

Wiggins, A., & Crowston, K. (2012). Goals and tasks: Two typologies of citizen science projects. In *2012 45th Hawaii International Conference on System Sciences* (pp. 3426–3435). IEEE. https://doi.org/10.1109/HICSS.2012.295

GOVERNING DEEPFAKES: LEGAL INITIATIVES AND REGULATORY GAPS

YASAMAN YOUSEFI,[1, 2]
MARIA DOLORES SANCHEZ GALERA,[3]
ANGELO TUMMINELLI,[4] CALOGERO CALTAGIRONE,[4]
TOMMASO TONELLO[5]

[1] DEXAI-Artificial Ethics, Rome, Italy
yasaman.yousefi@dexai.eu
[2] University of Bologna, CIRSFID ALMA AI, Faculty of Legal Studies, Bologna, Italy
y.yousefi@unibo.it
[3] Charles III University of Madrid, Madrid, Spain
mariadsa@inst.uc3m.es
[4] LUMSA University, Department of Human Sciences, Rome, Italy
a.tumminelli@lumsa.it, c.caltagirone@lumsa.it
[5] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
t.tonello@uu.nl

This chapter examines the pervasive threat of digital disinformation, with a specific focus on AI-generated content as a paradigmatic challenge to contemporary governance. The analysis blends ethical and legal perspectives to assess existing mitigation strategies. AIGC occupies a critical intersection of advanced technical capability, complex social meaning-making, and often conflicting legal protection frameworks. Consequently, effective responses require an interdisciplinary approach that integrates conceptual clarity, technical standards, robust legal instruments, and widespread social interventions to preserve public trust and protect vulnerable individuals.

DOI
https://doi.org/
10.18690/um.feri.2.2026.7

ISBN
978-961-299-109-8

Keywords:
Digital Disinformation,
EU Regulatory Approach,
GDPR,
AI Act
Harm Mitigation

University of Maribor Press

1 Conceptual Considerations

This section establishes the ethical and sociological context for disinformation, framing the problem of synthetic media in terms of relational responsibility and the material consequences of immaterial harms.

Conceptual clarity regarding the nature of digital communication is necessary to frame legal responses. Sociological critiques argue that the hyperconnected infosphere fosters a cultural state of "existential relativism," a condition where distinctions between truth and falsehood blur, rationality yields to emotionality, and communication operates under the premise that "anything goes" (Donati, 2024, p. 36). This phenomenon risks confusing technologies that support human identity with those that actively erode it, leaving individuals vulnerable to technological domination (Donati, 2024, p. 32).

The cultural diagnosis of "existential relativism" in techno-mediated contexts cannot remain a mere description of fragmented meanings. The pervasiveness of digital platforms destabilizes symbolic reference points and weakens shared norms. This sociological condition translates into normative challenges, requiring new forms of rule legitimation. At the same time, it generates moral challenges, expanding responsibility for actions whose consequences are diffuse. Subjectivity must therefore renegotiate criteria of autonomy and accountability. The shift toward ethical responsibility becomes a response to the volatility of digital environments. In sum, cultural diagnosis demands an ethical rethinking capable of guiding common practices.

In this sense, the concept of responsibility must be re-centred. Responsibility, in its deepest sense (Miano 2009; Da Re 2003), is not merely an individual legal commitment but a dialogical and ecological capacity to respond to the call of others and to care for the world as a shared home. The velocity and pervasive nature of AI challenge this relational commitment. The creation or sharing of deceptive content without reflecting on its impact constitutes a profound failure of this relational commitment.

When technological systems, such as hyperconnectivity and algorithmic amplification, overwhelm individual capacity for verification and responsible reflection, the individual alone cannot discharge the ethical duty of care. This creates an ethical vacuum. The regulatory response, namely, the requirement under the Digital Services Act (DSA) that Very Large Online Platforms (VLOPs) manage systemic risks, is thus ethically justified. The state enforces the transfer of the burden of relational care from the overwhelmed individual to the systemic actors (platforms) that control the informational infrastructure. However, is this enough? In addition, how can we trust self-regulation and self-risk-management systems?

1.1 Privacy, Reputation, and the Materiality of Immaterial Harms

AI-generated contents pose direct threats to protected rights, notably privacy and reputation, by weaponizing personal data.

- **Privacy:** Privacy is the inherent right of an individual to control their personal information, linked intrinsically to dignity, freedom, and autonomy. Deepfakes violate this right by depicting individuals in false, compromising, and potentially harmful situations without consent, attacking the integrity of their self-presentation.
- **Reputation:** Reputation reflects the moral and social value attributed to a person, based on actions and perceived identity, functioning as a critical component of credibility within a community. Deepfakes inflict grave damage by distorting public perception, leading to exclusion, professional loss, and emotional distress.

The Cambridge Analytica scandal illustrates how the misuse of personal data can become a powerful instrument of manipulation and reputational harm. By harvesting the personal information of millions of Facebook users without their knowledge or consent, Cambridge Analytica exploited intimate details of individuals' preferences, vulnerabilities, and networks to influence electoral behaviour (Isaak & Hanna, 2018). This case underscores how data, once weaponized, undermines privacy and autonomy by stripping individuals of control over their own digital identities, while simultaneously reshaping collective reputations and public discourse in ways that erode trust in democratic institutions.

Deepfakes exacerbate these concerns by combining the mass-scale data misuse seen in Cambridge Analytica with highly persuasive falsifications of identity. Unlike simple data profiling, deepfakes do not just predict or manipulate preferences; they fabricate "hyperreality". Comparable to revenge porn cases, where intimate images are shared without consent, or the proliferation of deepfake pornography targeting women in public life, these manipulations inflict enduring reputational damage that cannot be easily corrected once the falsified content circulates (Chesney & Citron, 2018). Similarly, instances where politicians or journalists are targeted with synthetic media, such as the 2019 deepfake video of Nancy Pelosi manipulated to make her appear intoxicated, demonstrate how fabricated content erodes public trust, polarizes societies, and destabilizes democratic debate (Reuters, 2020).

Critically, the harms inflicted by deepfakes are often immaterial: psychological distress, reputational degradation, and erosion of evidentiary trust. While these harms are not physical or pecuniary in the traditional sense, they carry severe material consequences (e.g., job loss, social ostracization). This profile presents a critical remedial gap. Current liability frameworks, including the revised Product Liability Directive (PLD), remain primarily oriented toward material or pecuniary damages, rendering the doctrinal fit for typical deepfake injuries imperfect and procedurally onerous for victims.

2 The Constitutional Balancing Exercise: Freedom of Expression, Human Rights, and Democratic Integrity

Effective mitigation strategies must navigate the tensions inherent in liberal constitutional orders, requiring a careful balance between freedom of expression and the protection of other fundamental rights, particularly the right to receive accurate information and the integrity of democratic processes. Accurate information and knowledge are necessary for citizens to make informed political decisions, as systematically deceitful content can distort the opinion-forming process, potentially leading to electoral results based on a perverted public discourse.

The challenge lies in reconciling these competing constitutional demands, a process heavily influenced by contrasting legal traditions across the Atlantic. The French approach illustrates these dilemmas vividly: the 2018 "fake news law" (Loi n° 2018-1202) empowers judges to order the removal of false or manipulated content, including deepfakes, during election periods if it is likely to affect the outcome of a vote. While designed to safeguard democratic integrity, the law has been criticized for its potential chilling effects on freedom of expression and the press, as the broad and somewhat vague definitions of "false information" risk overreach (Douek, 2025). Similar tensions arise across the EU, where regulation must remain consistent with the European Convention on Human Rights and the Charter of Fundamental Rights of the EU, both of which enshrine freedom of expression while also permitting proportionate restrictions necessary in a democratic society under the rule of law premises. This balancing act demonstrates that regulating synthetic media is a constitutional challenge as much as a technical one, requiring legislators and courts to calibrate carefully between the prevention of harm and the preservation of open discourse.

Freedom of expression in Europe, codified in Article 10 of the European Convention on Human Rights (ECHR) and Article 11 of the EU Charter of Fundamental Rights, is recognized as a *relative* right, not an absolute one. The European framework incorporates a crucial *passive dimension* of freedom: the right to receive information in a pluralistic context, explicitly linking it to the functioning of a "democratic society". European courts prioritize values such as human dignity and pluralism. Consequently, false, misleading, or deceitful information does not receive the unfettered constitutional protection afforded under the US model. The ECHR framework explicitly allows for limitations to freedom of expression when such limitations are deemed "necessary in a democratic society" (Article 10(2)). The European Court of Human Rights (ECtHR) has confirmed that the Internet environment poses a "higher risk of harm" compared to traditional media, justifying greater limitations, provided that the legislator provides the framework for reconciling competing claims. This distinction makes the European Union's resulting multi-instrumental regulatory stack (DSA, AI Act, GDPR) constitutionally permissible, as its foundation is the defence of the passive right to be informed and the preservation of pluralism against intentional disinformation. In electoral periods, freedom of political debate is paramount, but in cases of conflict, contracting states have a margin of appreciation to restrict speech to protect the "free expression of the opinion of the people in the choice of the legislature".

3 Regulatory Measures as Mitigation Strategies: The EU Architecture

The EU has developed a complex, multi-instrumental architecture, designed to govern AI and content dissemination across the entire lifecycle (design, deployment, dissemination, and remedy). These instruments operate as complementary levers, and introduce points of friction and structural limitations.

3.1 General Data Protection Regulation (GDPR): Friction, Accuracy, and the Technical Impracticability of Erasure

The General Data Protection Regulation (GDPR) is immediately relevant because deepfakes are frequently produced using personal data, including images or other associated information that can be traced back to an individual, such as someone's recognisable voice. Article 4(2) GDPR defines "processing" broadly, covering every stage from collection to dissemination, which clearly encompasses the creation and distribution of deepfakes. A key obligation here is the principle of accuracy under Article 5(1)(d), which requires controllers to take reasonable steps to ensure that inaccuracies in personal data do not cause harm. Generative models that produce fabricated likenesses or statements implicate this principle when the output is traceably linked to an identifiable individual, particularly where reputational or dignitary harm follows.

Supervisory authorities have already begun to test the GDPR's applicability in this context. In 2022, the Italian Data Protection Authority (*Garante*) launched an investigation into *FakeYou*, a platform offering synthetic voice generation of public figures, to determine how personal data were being processed and whether safeguards against misuse were in place (Garante per la protezione dei dati personalo, 2022). More recently, in October 2023, the *Garante* adopted an urgent measure against *Clothoff*, an app that generated "deep nudes" by creating pornographic content from images of real people. The authority imposed the immediate limitation of data processing for Italian users, stressing that the service allowed anyone, including minors, to create synthetic sexualized content without verifying consent and without any indication of the artificial nature of the images. These cases show that EU data protection authorities view the misuse of deepfake technologies as a clear form of unlawful processing under the GDPR, particularly when fundamental rights such as dignity, privacy, and the protection of minors are at stake (Garante per la protezione dei dati personali, 2025*)* .

Despite this, enforcement faces significant technical friction. The right to erasure (Article 17) illustrates the problem: even if a data subject requests deletion, trained AI models may retain informational traces that allow re-synthesis of a likeness. This raises the need for controllers to ensure lawful data provenance and consent before training occurs, as post hoc deletion is technically challenging if not impossible. Further complexity arises from contextual exemptions, such as the household exemption (Recital 18), which can shield the private creation of harmful deepfakes from GDPR scrutiny until dissemination occurs, creating a regulatory gap at the point of initial harm generation.

Ultimately, effective governance of deepfakes depends on aligning controller obligations under GDPR with the transparency and traceability requirements mandated by the forthcoming AI Act. Without rigorous enforcement of data provenance and consent under GDPR, subsequent interventions under the Digital Services Act (DSA) and AI Act risk becoming reactive, addressing harm only after it has occurred rather than preventing it at the source.

3.2 EU Artificial Intelligence Act (AI Act): The Limited-Risk Paradox and the Transparency Regime

The artificial intelligence Act (AI Act Regulation (EU) 2024/1689), the world's first comprehensive legal framework on AI, represents the EU's most explicit statutory engagement with synthetic media. The AI Act provides a legal definition of deepfakes: "AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful" (Art. 3(60)).

The AI Act situates the problem of deepfakes within a political and ethical frame by foregrounding the risk of manipulation. Recitals 28 and 29 explicitly identify deception and manipulation among the principal social risks arising from the misuse of generative technologies, warning that such misuse can impair democratic processes and corrode public trust. Recital 133 further reiterates the legislative purpose of enabling individual recipients to recognise synthetic content and guard against impersonation and deceit.

The AI Act employs a risk-based approach, which includes a hard prohibition under Article 5 for AI systems categorized as posing an unacceptable risk. Specifically, Article 5 prohibits AI systems that use subliminal techniques or manipulative or deceptive techniques to distort behaviour, potentially causing physical or psychological harm. It also prohibits systems that exploit the vulnerabilities of individuals or specific groups. This provision sets a critical boundary against the most dangerous forms of manipulation.

For the vast majority of deepfakes, the AI Act addresses them through a mandatory transparency regime anchored in Article 50. This article imposes a dual obligation: providers of generative systems must ensure that outputs are marked in a machine-readable way, and deployers who disseminate synthetic content must disclose to the public that the material has been generated or manipulated. This infrastructure aims to make provenance and traceability foundational elements of the digital information ecosystem.

However, deepfakes are classified primarily as a *limited-risk* category, thereby avoiding the stringent substantive and supervisory requirements imposed on high-risk systems. This policy choice, intended to protect innovation and legitimate expressive uses, risks significant under-protection in contexts where manipulation yields acute public-interest harms, such as targeted electoral interference. The Act's reliance on transparency is vulnerable to adversarial evasion, as malicious actors can deliberately strip metadata or disseminate content via decentralized channels, thereby nullifying the prophylactic intent of Article 50. Moreover, the disclosure duty, linked to the standard of the "reasonably well-informed, observant and circumspect user", risks implicitly burdening less media-literate populations with verification duties, attenuating protection for those most susceptible to manipulation.

The AI Act's reliance on transparency is thus recognized as necessary but not sufficient to counter sophisticated manipulation, particularly in high-stakes political contexts where systemic democratic harm is the risk.

3.3 Digital Services Act (DSA): Reactive Moderation, Systemic Risk, and Enforcement Gaps

The Digital Services Act (DSA) is central to content governance, placing distinct obligations upon online intermediaries for content moderation, transparency, and, crucially, systemic risk assessments. For Very Large Online Platforms (VLOPs), the DSA mandates the identification and mitigation of systemic risks, including those arising from disinformation and algorithmic amplification.

Despite its importance, the DSA's efficacy is constrained by several limitations. First, its mechanisms are largely *reactive*, operating through notice-and-action procedures after content has already been posted. While effective in mitigating ongoing harm, reactive measures cannot restore eroded public trust or undo immediate reputational injury. Second, the DSA focuses primarily on large, regulated platforms, neglecting important vectors of dissemination such as decentralized protocols and private messaging applications frequently used to circulate deepfakes. Third, enforcement relies on platform cooperation and transparency. Compliance monitoring, particularly concerning soft-law commitments like the Code of Practice on Disinformation, has been assessed as uneven and often lacking methodology-opaque reporting (Böswald, 2025). Therefore, while the DSA complements the AI Act by addressing dissemination, it does not negate the need for proactive provenance and detection at the generation point.

3.4 Product Liability Directive (PLD)

It is also important to note that the Product Liability Directive (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, the 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, the PLD provides only partial remedies and relies heavily on the AI Act to fill liability gaps. As scholars note, further legislative refinement will be necessary to extend liability to the full spectrum of harms typically caused by generative AI, especially in cases where reputational damage and privacy violations constitute the primary injury.

3.5 Soft Law Mechanisms

Legally binding regulation plays an important role in combating AI-generated disinformation. Nonetheless, policy research and policy negotiation efforts for the legislative process represent time-consuming activities (Schepel, 2005). Soft law tools in the form of non-binding norms, guidelines, codes of practice, and so on, can help manage lengthy regulatory processes by encouraging voluntary compliance from different stakeholders. Considering the fast-paced innovation in the generative AI context, earning it a place among disruptive technologies, soft law instruments promote flexible and timely reactions to promote ethical AI governance and to collect information on the empirical effects of soft law compliance (Păvăloaia & Necula, 2023).

In a context of international diplomatic and economic tension, however, it is argued that the non-binding nature of soft law raises concerns over its ability to attract stakeholders and encourage their compliance, contributing to concerns of a crisis of global AI governance (Leslie & Perini, 2024). The risk highlighted by the two authors is more real for some than others. The EU is particularly exposed to the flaws of soft law in the AI race: since 2001, the Commission has expressed interest in externalising governance duties by fostering the involvement of private stakeholders in contributing to relevant policy through self- and co-regulatory, i.e., non-binding measures.

Additionally, the legal challenges of AI appear particularly urgent considering the Union's role as a normative power: while the EU has traditionally leveraged on his large internal market to foster international companies" adaptation to European legal standards, including in the context of the fight to online disinformation, geopolitical attrition seems to undermine the principle of voluntary compliance that makes soft law a helpful tool in protecting online information and digital citizens' rights (Manners, 2002). By stressing soft law's complementary role *vis à vis* legally binding regulation, this section argues that integration of soft law tools in hard law covenants may foster AI regulation and, more specifically, the fight against AI-generated disinformation and deepfakes.

The recent endorsement of the European Commission and of the European Board for Digital Services of the 2022 Strengthened Code of Conduct in the Digital Services Act (DSA) points in this direction. The goal of contextualising the Code of

Conduct within EU regulation is allegedly to ensure better compliance with EU law for AI service providers and, consequently, to clearly define accountability.

The persistence of an accountability gap is motivated by several factors, some of them already been referenced earlier. In the first place, there persists a struggle to regulate AI, which is in turn related to both economic competitiveness concerns and to the technology-induced legislative lag (European Commission, 2025; Kosta et al., 2025). On the other hand, issues related to the opacity of AI algorithms and to our ability to attribute agency, and therefore, accountability, to AI algorithms hinders the legislator's ability to "show that the issues have been conscientiously addressed and how the result has been reached; or alternatively alert the recipient to a justiciable flaw in the process"(Calderonio 2025; Floridi 2023; Williams et al. 2022).

A great deal of uncertainty in relation to accountability, moreover, stems from the semantic uncertainty surrounding the concept, given its fluid, i.e., context and discipline-dependent, meaning. Williams et al. suggest that abstract aspirations, such as the principle of accountability, need to be specific and enforceable, an applicability gap also highlighted by Leslie and Perini. They argue that, by mapping the semantic debate on accountability, it is possible to identify five concepts, related chronologically in terms of how these terms are related, which inform the others, and how, as well as from an "activity" perspective. By the latter, it is meant how these terms foster push-pull dynamics or, in other words, to clarify whether AI providers are required to make information available (push) or if it is end-users who seek information in each context (pull). According to the authors, accountability is the last step necessary to make the concepts listed above enforceable. At the same time, these principles allow for framing accountability differently depending on the (AI) system under inquiry, making these aspirations capable of being enforced and of managing different AI systems.

It becomes then clear that frameworks like the one proposed by Williams et al. represent a necessary step to move from principles to practice, even if further challenges posed by generative AI to delineating AI agency and accountability will require a fine-tuning of such models. Nonetheless, integrating soft law instruments against disinformation in legally binding documents represents an attempt to bolster the commitment to the fight against disinformation, as well as a necessary step to deliver the tools and the metrics to tackle the AI services providers' accountability gap.

Now, soft law tools such as the Strengthened Code of Practice envisage objectives for signatories such as the following: the release of periodic transparency reports covering volumes of synthetic content, the number, and outcomes of reports and takedowns; the use of standardized reporting templates to enable comparative evaluation; cooperation during sensitive events, e.g., electoral periods. However, cooperation from this perspective has at times been sluggish, with AI companies providing limited and incomplete information or sloppy justification for the data collection methodology that they presented (OECD, 2024).

By contextualising such soft law tools into legally binding regulation (such as the DSA), nonetheless, it would be possible to frame accountability issues within a specific policy setting. If, on the one hand, this would eventually prompt EU institutions to defend their reliance on codes of conduct *et similia* in the AI governance context, on the other hand, it also articulates those push factors that AI service providers need to be presented with, as advocated for by Williams et al.

In short, soft law tools represent an important means to foster the objectives of documents that articulate compulsory actions, such as the DSA. By being contextualized within binding documents, it becomes possible to move from ethical AI governance principles advocated for in soft law tools to their practice.

4 Structural Challenges in the Governance of Deepfakes

Despite the EU's increasingly dense regulatory ecosystem, deepfakes expose persistent structural vulnerabilities in law's capacity to safeguard democratic integrity and individual dignity. The problem is not merely the presence of malicious actors but the systemic asymmetries between rapid technological development and the slower pace of legislative adaptation, the uneven enforcement capacities across Member States, and the incomplete coverage of harms, particularly immaterial and distributive ones. This section identifies five interlinked shortcomings in the current governance framework.

1) **Technological-Legislative Asymmetry:** the foundational challenge is the inherent disparity in speed between technological innovation and regulatory response. Generative capabilities evolve rapidly, meaning detection techniques (such as inference-based methods) and provenance architectures (such as watermarking) are often one step behind.

The AI Act's reliance on transparency is vulnerable to adversarial evasion strategies. Malicious actors can deliberately strip metadata, transcode files, re-edit labelled outputs, or employ adversarial attacks to obfuscate generation signatures, effectively nullifying the prophylactic intent of Article 50. The AI Act requires to track the provenance of AI-generated media. However, it does so without requiring sustained public investments towards detection research risks. At the same time, it delegates enforcement to private stakeholders, who may lack the necessary resources or incentives. A resilient architecture requires proactive measures, including public funding for detection research and the standardization of robust, tamper-resistant provenance mechanisms that prioritize interoperability.

2) **The Honest-Actor Problem and Transnational Enforcement Deficits:** EU legal instruments principally regulate actors with a clear EU nexus, providers, deployers, and platforms operating in the Union. However, deepfakes are easily disseminated across borders, and malicious actors often operate from jurisdictions with weak enforcement capacity or through highly decentralized protocols.

The ease of cross-border dissemination enables sophisticated evasion strategies. This governance gap means that domestic legal obligations risk producing mere "protective islands" that are porous at their boundaries. Addressing this honest-actor problem requires robust international cooperation, harmonized standards for provenance and liability, and the establishment of reliable bilateral and multilateral channels for rapid content takedown and mutual legal assistance.

3) **Fragmentation and Enforcement Deficit:** The multi-instrumental nature of EU regulation, involving the AI Act, DSA, GDPR, and PLD, creates both overlap and complexity. While redundancy can increase robustness, complexity undermines clarity for regulated entities. Divergent interpretations of obligations by various enforcement bodies, national data protection authorities, digital services coordinators, and national courts exacerbate this issue.

Furthermore, uneven enforcement capacity across Member States results in varied levels of protection. This fragmentation risks creating a 'forum-shopping' environment, whereby bad actors may seek legal solutions in more permissive countries and, by doing so, hindering seeking remedies face and increasing procedural hurdles and uncertainty for victims. Uniformity in enforcement is necessary to ensure the protective effect of the EU stack is realized consistently.

4) **Insufficient Coverage of Immaterial Harms:** A core normative lacuna remains the treatment of immaterial harms. Deepfakes frequently inflict severe non-material injuries, including the erosion of dignity, emotional distress, and reputational degradation. Existing liability frameworks, such as the PLD, are slowly adapting to AI but remain historically oriented toward material or pecuniary damages.

   To bridge this remedial gap, substantive legal reform must be complemented by procedural innovation. Because the velocity of harm propagation is high in the digital sphere, temporal responsiveness is critical. Regulatory design should incorporate expedited administrative remediation pathways, such as mandates for swift provisional injunctive relief or statutory entitlements to rapid removal of non-consensual or clearly falsified content, in addition to traditional civil damages.

5) **Uneven Impact and Distributive Vulnerability:** Deepfakes do not affect all populations equally. Empirical evidence indicates a clear differential impact, with victims of non-consensual intimate imagery overwhelmingly being women, and marginalized groups frequently targeted by political disinformation campaigns (Kira, 2024). A regulatory architecture prioritizing technological neutrality may unintentionally fail to centre distributive justice and dignity.

   Addressing differential impact requires a rights-sensitive lens in regulatory design. Legal and policy responses must prioritize protective measures for the most vulnerable groups, ensuring mechanisms such as expedited takedown are readily accessible, alongside legal aid and psychosocial support. Furthermore, platform operators and regulators must incorporate explicit distributive impact assessments as part of their systemic risk frameworks (under the DSA), ensuring that mitigation efforts do not merely shift harms to less visible spaces or less empowered communities.

5 Concluding Remarks

In this chapter, we first framed the concept of disinformation through a socio-ethical lens, looking at how relational responsibility and the material consequences of immaterial harms emerge from the "hyperconnected infosphere". Second, we highlighted different traditions in the freedom of expression and its legal understanding, in order to justify effective mitigation strategies and their reach. Third, we presented the EU's complex AI governance architecture and analysed the legal instruments on which it leverages. Fourth, we identified five shortcomings hindering EU AI governance.

The analysis of mitigation strategies demonstrates that the EU's governance architecture for synthetic media is characterized by groundbreaking, innovative intent and significant structural insufficiencies. The classification of deepfakes as limited-risk under the AI Act, combined with the inherently reactive nature of platform governance under the DSA, limits the overall efficacy of the regulatory stack.

The evidence from constitutional traditions and case studies confirms that disinformation, particularly when amplified by synthetic media, constitutes a systemic threat to democratic participation. Effective governance must be rooted firmly in the European constitutional commitment to *pluralism* and the passive right to *informed choice*, justifying intervention that transcends the US marketplace paradigm.

However, persistent challenges, the technological-legislative asymmetry, the honest-actor problem, and the insufficient legal coverage of non-material harms demand strategic recalibration. An effective, future-proof governance strategy requires a coordinated policy shift that moves resolutely beyond a transparency-only paradigm for high-consequence contexts. The priorities must include substantial public investment in interoperable detection and tamper-resistant provenance standards (as technical solutions complement legal frameworks), securing international regulatory harmonization, and creating procedural mechanisms tailored to expedite remedial action for reputational and psychological harms. These diagnostic conclusions form the essential analytical groundwork for the integrated policy recommendations that will be detailed in Chapter 8.

End notes

Yasaman Yousefi is the lead author of this Chapter. She conceptualized this Chapter, coordinated the writing and polished the final text. She wrote the introduction and the conclusion, as well as the following sections: The Constitutional Balancing Exercise: Freedom of Expression, Human Rights, and Democratic Integrity, Regulatory Measures as Mitigation Strategies: The EU Architecture, and Structural Challenges in the Governance of Deepfakes. Maria Dolores Sanchez Galera contributed to the legal analysis in Regulatory Measures as Mitigation Strategies: The EU Architecture. Angelo Tumminelli and Calogero Caltagirone wrote the conceptual considerations. Tomasso Tonello wrote the Soft Law Mechanisms. All authors reviewed and approved the final version.

References

Böswald, L.-M. (2025, November 19). Soft law, hard risks? Co-regulation and risk mitigation under the Digital Services Act. Interface. https://www.interface-eu.org/publications/dsa-co-regulatory-mechanisms

Calderonio, V. (2025). *The opaque law of artificial intelligence* (arXiv:2310.13192). *arXiv*. https://doi.org/10.48550/arXiv.2310.13192

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

Da Re, A. (2003). *Filosofia morale: Storia, teorie, argomenti*. Pearson Italia.

Deep fake: Garante avvia istruttoria su app che falsifica le voci. (2022, October 12). *Garante per la protezione dei dati personali*. https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9816291

Deepfake, Garante privacy: Stop a Clothoff, l''app che spoglia le persone. (2025, October 3). *Garante per la protezione dei dati personali*. https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10174320

Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC (Text with EEA relevance). (2024). *Official Journal of the European Union*. http://data.europa.eu/eli/dir/2024/2853/oj

Donati, P. (2024). *Being human in a virtual society*. Peter Lang. https://www.peterlang.com/document/1461639

Douek, E. (2024). The Politics and Perverse Effects of the Fight Against Online Medical Misinformation. *Yale LJF*, *134*, 237.

European Commission. (2025, February 13). *Commission endorses the integration of the voluntary Code of Practice on Disinformation into the Digital Services Act*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/commission-endorses-integration-voluntary-code-practice-disinformation-digital-services-act

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*(1), Article 15. https://doi.org/10.1007/s13347-023-00621-y

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer, 51*(8), 56–59. https://doi.org/10.1109/MC.2018.3191268

Kira, B. (2024). *Deepfakes, the weaponisation of AI against women and possible solutions*. *Verfassungsblog*. https://doi.org/10.59704/9987d92e2c183c7f

Kosta, E., Hallinan, D., Hert, P. D., & Nusselder, S. (2025). *Data protection, privacy and artificial intelligence* (Vol. 17). Bloomsbury Publishing.

Leslie, D., & Perini, A. M. (2024). Future Shock: Generative AI and the international AI policy and governance crisis. Harvard Data Science Review, (Special Issue 5).

Manners, I. (2001). Normative Power Europe. A contradiction in terms, 235-258.

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.

OECD. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/facts-not-fakes-tackling-disinformation-strengthening-information-integrity_ff96d19f/d909ff7a-en.pdf

Reuters. (2020, August 3). *Fact check: "Drunk" Nancy Pelosi video is manipulated*. https://www.reuters.com/article/world/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2B1/

Schepel, H. (2005). The constitution of private governance: Product standards in the regulation of integrating markets (Vol. 4). Hart Publishing.

Păvăloaia, V. D., & Necula, S. C. (2023). Artificial intelligence as a disruptive technology – a systematic literature review. Electronics, 12(5), 1102.

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

Regulatory Innovations and Policy Options for Synthetic Media and Digital Democracy

Andrew McIntyre,[1] Yasaman Yousefi,[2, 3]
Maria Dolores Sánchez Galera[4]

[1] University of Amsterdam, Amsterdam, the Netherlands
a.mcintyre@uva.nl
[2] DEXAI-Artificial Ethics, Rome, Italy
yasaman.youefi@dexai.eu
[3] University of Bologna, CIRSFID ALMA AI, Faculty of Legal Studies, Bologna, Italy
y.yousefi@unibo.it
[4] Charles III University of Madrid, Madrid, Spain
mariadsa@inst.uc3m.es

DOI
https://doi.org/
10.18690/um.feri.2.2026.8

ISBN
978-961-299-109-8

This chapter explores potential regulatory innovations and policy options for addressing the democratic risks and opportunities of AI-generated content (AIGC) within the European context. Drawing upon and responding to discussions in previous chapters, it argues that current policy approaches centred on the detection, moderation and containment of AIGC are not only insufficient but also risk reinforcing authoritarian tendencies. Instead, the chapter outlines a policy strategy that emphasizes political participation and pluralism as a means of promoting democratic resilience and addressing the specific harms of AIGC. This strategy is oriented around three key objectives: (i) clarifying AIGC harms, (ii) strengthening institutional coordination, and (iii) enhancing digital literacy and citizenship. Key to this strategy is the reconceptualization of generative AI as a creative and expressive tool for promoting more inclusive political dialogue and democratic debate. Ultimately, this chapter envisions a future in which GenAI is not solely understood as a threat to democracy but as a resource for fostering a more trustworthy information environment and political system. It is a future where truth may become increasingly difficult to determine, but in which our democratic values nonetheless remain protected and strengthened.

Keywords:
regulatory innovation,
counter-disinformation
policy recommendation,
pluralism,
digital democracy,
political participation

University of Maribor Press

1 Policy and pluralism

Building on the analysis of democratic risks in Chapter 5 and critiques of mitigation strategies in Chapter 6, this final chapter examines how harmful AI-generated content (AIGC) is conceptualised in current European policy and proposes new governance strategies. To begin, section 8.1 explores the unique challenges of counter-disinformation policy, showing how measures aimed at governing truth may erode democratic trust and promote authoritarian tendencies, highlighting the need for active citizenry and pluralist debate. Beyond addressing the negative impacts of AIGC, section 8.2 then considers how GenAI could be utilised as a unique tool of representation and communication that can promote pluralist debate and political participation. Finally, section 8.3 builds on these discussions to outline priority areas for policy as part of a broader strategy that addresses harms while promoting democratic resilience. This requires clarifying harms, acknowledging tensions, and reconceptualising AIGC as socio-political resources rather than solely risks that need to be mitigated.

Before discussing European policy specifically, it is necessary to briefly frame this policy discussion within the broader epistemic context of GenAI. As Floridi argues, we now exist in an infosphere where human experience and knowledge are redefined in terms of information flows (Floridi, 2014). From this perspective, AIGC does not simply mislead individuals; it contributes to and alters the structural integrity of our wider information environment (Russo, 2022). Beyond introducing artificial content, AIGC reshapes the epistemic conditions under which societies construct, verify, and contest knowledge (Bisconti et al., 2024). Disruption has profound implications for collective knowledge, socio-political discourse, and democratic deliberation (McIntyre et al., 2025). AIGC is not inherently detrimental, but its use for disinformation presents what we describe as *informational harms*.

As Feinberg argues, harm is a wrongful infringement or obstruction of a person's interests. These interests include one's physical safety and further extend to other interests such as property, privacy, autonomy, and reputation, among others. Therefore, harm can be both tangible (e.g., physical violence, theft) and intangible (e.g., violating privacy, restricting autonomy) (Feinberg, 1987). Within Floridi's infosphere, however, human beings are redefined as informational organisms whose identity, agency, and interests are fundamentally constituted by information flows and structures within our broader informational environment. Through this theoretical lens, we reconceptualise Feinberg's notion of harm as an infringement or obstruction of a person's informational integrity. As a person's informational being is embedded within and continually shaped by the wider infosphere, however, protecting individuals from harm ultimately depends on maintaining the integrity of the information environment as a whole. Thus, informational harms relate to how people are impacted by deception, misrepresentation, and disinformation, and how processes of knowledge construction, dissemination, and reception are impacted by the social integration of AI systems and the widespread production of AIGC.

To translate the notion of informational harms into policy, we draw on Smuha's harm categories related to AI. As Smuha argues, harms can be categorised at 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) (Smuha, 2021). For example, the 2024 US presidential election, marked by a surge in AIGC, exemplifies societal harms by eroding trust in institutions. The EU recognizes such risks in the AI Act, which acknowledges GenAI may generate material or immaterial harm (European Union, 2024). Yet existing frameworks remain reactive, focusing on moderation and detection rather than systemic impacts.

This chapter outlines policy priorities that address harms across these different levels while grappling with tensions such as institutional dysfunction and reconciling regulation with freedom of expression. Confronting these directly, the chapter offers a blueprint for reconceptualising AIGC as a potential resource for democratic resilience.

The European legal mechanisms discussed in Chapter 7 offer only limited solutions to the significant challenges posed by harmful AIGC. Many of these mechanisms are narrow in scope and practical application, failing to fully account for the deep integration and diverse use of GenAI in everyday life. As such, these frameworks do not adequately define or conceptualise AIGC as a socio-political phenomenon, nor do they address the diverse harms that AIGC can inflict upon different levels of society (individual, collective, societal). In section 8.3, we elaborate on possible legal innovations to more appropriately address the harms associated with AIGC as part of our wider policy priorities. However, legal solutions alone cannot fully account for the deep social integration and diverse use of GenAI in everyday life. As such, we need more diverse policy interventions and strategies for combating the spread and impact of harmful AIGC, as well as solutions for promoting stronger democracies.

Broadly speaking, emerging policy strategies fall into one of three categories: (i) retreat strategies aimed at reducing digital interactions in favour of in-person interactions to improve trust relationships; (ii) containment strategies aimed at detecting, labelling and limiting the impact of harmful AIGC; and (iii) mobilization strategies aimed at harnessing GenAI to promote more robust democratic systems (Allen & Weyl, 2024). Largely, states have pursued containment strategies as they focus on practical and tangible technological, legal, and social solutions and allow for the strict regulation of harmful AIGC. However, though well-intentioned in their attempt to protect informational integrity and democratic stability, many of these containment strategies seek to re-establish a single authoritative source of truth and, in doing so, paradoxically undermine democracy while reinforcing anti-democratic tendencies. To elaborate, let us critically examine the goals and assumptions underpinning these strategies, which Farkas and Schou divide into four dimensions: (i) policing the truth; (ii) re-establishing centres of truth-making; (iii) promoting public immunity; and (iv) technological solutionism (Farkas & Schou, 2023).

To elaborate, many containment strategies are aimed at policing truth, often relying on restrictive legislation and other drastic measures that policymakers justify as protecting the democratic foundations of truth and reason. However, Farkas and Schou describe such measures as authoritarian in that they are veiled attempts at censorship that consolidate government control over the information environment. Furthermore, these strategies shift open political debate into closed governmental mechanisms, which are rarely subject to public scrutiny. Secondly, often these efforts aim to re-establish traditional centres of truth-making (e.g., politics, science, journalism) and position these institutions as vital protectors of truth that must reclaim authority. Science, in particular, is often privileged above others, with researchers and technologists arguing that they should be included in high-level decision-making, even to the point of superseding public opinion. However, Farkas and Schou claim that these approaches risk emboldening certain groups as arbiters of truth, reinforcing the elitist notion that governance should be dictated by technocratic experts rather than public dialogue. Similarly, public education initiatives (e.g., media literacy programmes) aimed at strengthening individual critical thinking are certainly important and beneficial. However, these strategies are often framed as a method of curing public ignorance or immunising the public against manipulation. Farkas and Schou argue that such a framing places responsibility on individuals rather than governments or technology companies, while also dismissing popular dissent and diverse opinions as ignorance or delusion that is simply wrong in comparison to the single truth defined by experts.

Such strategies also often utilise advanced technologies, including AI systems, in order to detect, verify, and manage disinformation. While certainly technical innovations can be effective and beneficial, often these technical fixes are presented as the only viable solution and are too simplistic to fully address nuanced socio-political challenges. Furthermore, this relies upon private technology companies and gives these companies control over what constitutes truth and societal harm (Allen & Weyl, 2024).

This is not to say that technological solutions are inherently problematic and, indeed, we advocate for the ethical and transparent use of AI systems below. However, we wish to highlight that the blunt use of technologies to determine truth and harm risks undermining democracy further.

While we largely agree with Farkas and Schou's critiques and agree that we should not be attempting to arbitrate truth, we would not fully condemn or abandon these containment strategies.

These strategies offer partial solutions, but in the rush to combat disinformation, they may inadvertently undermine the very democratic values they seek to protect. The challenge, therefore, is not to discard these policies altogether but, rather, to implement them with a heightened awareness of the risks and ensure that they are designed to promote a more resilient, rather than a more controlled democracy.

This approach forms the core of the policy priorities presented in section 8.3 of this chapter. However, we must go further than simply careful and ethical implementation of containment strategies that seek to determine and arbitrate truth. As Farkas and Schou argue, we require an alternative approach for strengthening democracy that is not about establishing a single truth at all. Instead, they advocate for a pluralistic and genuinely political public sphere that embraces the "always-antagonistic dimension of the political" by fostering "spaces for vibrant clashes of conflicting alternatives" (Farkas & Schou, 2023).

Drawing on the work of political philosophers like Chantal Mouffe (Mouffe, 1997), Ernesto Laclau (Laclau, 1990), and Jacques Rancière (Rancière, 2014), Farkas and Schou contend that the current post-truth political crisis is not due to a lack of facts or an increase in deceptive media. Instead, it stems from a lack of meaningful democratic participation. More specifically, they argue that a healthy democracy is not about reaching a rational consensus on what is true but, rather, about embracing a culture of constructive and agonistic pluralism that involves a vibrant clash of democratic political positions. Therefore, instead of focusing solely on counter-disinformation measures, Farkas and Schou argue that policymakers should couple these measures with strategies that encourage greater and more diverse political participation; "more politics" rather than "more truth".

With the arrival of GenAI, we are fast approaching a world in which everyday people, not only states and companies, are powerful media producers capable of creating and distributing convincing AIGC around the world in moments. In such a world, retreat strategies are impractical and potentially detrimental in that technology bans are unlikely to be adopted by states, and it is unrealistic to expect people to voluntarily abandon digital life.

Even if this were achieved, they risk undermining the positive political uses of digital technologies (e.g., increased communication and representation), while squandering further potential uses of GenAI. Furthermore, containment strategies can only go so far and risk fostering authoritarian tendencies and exacerbating distrust in democratic institutions, as discussed. If we accept that the proliferation and social integration of GenAI will continue at pace, we cannot solely rely on retreat or containment strategies. Instead, it is necessary to embrace mobilization strategies that utilise GenAI to promote political engagement and agonistic pluralism. Where Allen and Weyl highlight the use of such systems for authentication, data privacy, and promoting access to public information spaces, we contend that AIGC can play a role in this constructive agonistic dialogue and could be used to promote democratic resilience.

2 Deepfakes for political participation

Much attention has been paid to the negative impacts of AIGC, and rightly so, given their origins in deepfake pornography and the imminent threats they pose to democracy. Not only does AIGC risk misrepresenting the actions and statements of individuals, but it also impacts the integrity of our information environment and disrupts communication between citizens or groups of citizens, thus undermining democratic processes of collective decision-making. As Mathias Risse argues, for citizens to make collective decisions on policies and laws that will affect the population, they require "a decent level of knowledge about the people with whom they share a polity, lest these citizens be deceived, e.g., about how certain measures affect others or what such people's worries are (Risse, 2023). Harmful or deceptive AIGC may lead to greater misunderstandings or animosity between different communities, encouraging political polarization that stifles collaboration and dialogue. However, there are more diverse uses of AIGC that have received less public attention but that indicate how GenAI could be utilised to promote democratic values and political participation.

This discussion focuses on those instances in which AI-generated content has been used to improve public engagement with socio-political discourse and/or encourage communication and empathetic connection between citizens. These instances might include, for example, translating government communications to engage with multi-lingual communities (e.g., Manoj Tiwari speaking Haryanvi in 2020 (Jee, 2020)), creating interactive education tools or exhibitions to better explain historical events and figures (e.g., *Dalí Lives* exhibition (Lee, 2019)), or visualising future scenarios to better communicate the consequences of abstract policy issues (e.g., *This Climate Does Not Exist* (Tousignant, 2021).

A particularly illustrative example is the exhibition *EXHIBIT A-i* (Blackburn 2023), which used GenAI to visualise the witness statements of 32 refugees previously held at Australia's offshore detention centres on Manus Island and Nauru (Doherty, 2023). Gathered by the law firm Maurice Blackburn, these witness statements explained in graphic detail the inhumane conditions of these centres and the regular incidents of violence, abuse, self-mutilation, rape, and suicide that occurred there. As reporters were restricted from accessing these centres, no photographs or recordings exist, and so a text-to-image GenAI system was used to produce visual representations. It is important to note that these synthetic images were not intended as deception or as a substitute for evidence and their artificiality is openly acknowledged in the exhibition. Regardless, these artificial images provide the public with a bleak and visceral depiction of life in these centres and thus enable a more intimate understanding of the experiences of real people than can be achieved through text alone. Such images emphasize the human and personal impact of immigration policies, thus allowing citizens to better assess the actions of government institutions and the choices made by those politicians and officials in positions of power.

While these positive uses of AIGC are currently rare and often regarded as little more than curiosities or artistic experiments, they highlight the potential of how GenAI might be used to improve socio-political participation and epistemic agency. With greater and more engaging access to information about historical events, other communities, and the real and potential impacts of said policies on different communities, citizens may be able to more effectively formulate their own political opinions, empowering them to more competently engage with political discussions and to more confidently exercise their political agency in collective decision-making processes.

In Chapter 6, we explored the use of AI-generated content to promote specific values that aligned with the United Nations Sustainable Development Goals (SDGs). While they offer a creative and engaging way of communicating the SDGs, many participants in our use case expressed concern about the potential for deception and political manipulation, as well as the ethics of using historical or deceased figures to promote certain ideas without consent. These concerns echo those of Farkas and Schou with regard to authoritarian tendencies and the policing of truth. Rather than utilise GenAI to communicate selected values perceived as democratic (e.g., SDGs), it seems more appropriate and more democratic to place these technologies in the hands of citizens themselves and to encourage ethical use in public communication. As this technology becomes more deeply embedded into our everyday lives and communicative practices it has the potential to strengthen pluralist debate and remove barriers to political participation.

Previously, a lack of resources (e.g., finances, time, technology) or limited communicative capabilities (e.g., storytelling, oratory, technical skills) might have restricted citizens from fully participating in democratic dialogue and decision-making. With GenAI more widely available, however, the average citizen needs only provide a simple prompt to rapidly produce expressive, empathetic, and engaging audiovisual content representing their daily life. In doing so, individuals could easily visualise their personal experiences and private events that might otherwise go undocumented or ignored. This could include instances of systemic violence, abuse, and neglect, ensuring that the injustices and inequalities that citizens endure are visualised in detail, in ways that resonate with the wider public.

This is not to argue for a purely technological solution but rather to highlight how such technologies might be utilised through mobilization strategies to promote democratic values. Certainly, the widespread use of GenAI has significant risks (e.g., pornographic abuse, disinformation), but if appropriately implemented, this technology could enable citizens to better appreciate the lives of other communities, to engage with a plurality of views, and to understand how government policies and legislation might impact one another differently. Recalling Farkas and Schou's constructive antagonism, the purpose of such strategies is not to arbitrate truth but, rather, to promote a more vibrant, creative, and plural political debate. Coupled with light-touch containment strategies and legislative innovations, we may begin to move toward a more trustworthy information environment and political system wherein truth may become increasingly difficult to ascertain but wherein our democratic values are nonetheless upheld. The use of AIGC for promoting political engagement, alongside containment and literacy strategies, forms a key aspect of our proposed policy priorities described in the next section.

3 Regulatory and policy priorities for democratic resilience

Based on the above discussion, we propose that a strategy for democratic resilience should be aimed at maintaining the integrity of our information environment and, rather than arbitrating the truth, promoting a technically literate and politically active citizenry. While we recognise the need for containment strategies and technological solutions, this strategy emphasizes societal adaptation through conceptual unity in law and policy, robust democratic systems, and social integration of AI. This strategy builds upon the specific measures recommended by the European Parliamentary

Research Service (EPRS), as well as other existing counter-disinformation policy and regulatory proposals. It aims to address harms across Smuha's three levels of harm (individual, collective, societal) and is oriented around three key objectives: (i) legal clarification of AIGC and informational harms; (ii) coordination of democratic institutions; and (iii) promoting plural and participatory citizenship. These priority proposals are explained in more detail below, while Table 8.1 illustrates how they are aligned with the strategic objectives and how they address the levels of harm.

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)

Source: Own

3.1 Unified Legal Framework on Synthetic Media

Across European legislation, policy, and counter-disinformation strategies, the specific issue of AIGC is ill-defined. In the context of AI governance legislation and policy (e.g., AI Act, national AI strategies), the harms of AIGC are noted as a concern, but other socio-political issues (e.g., algorithmic bias, surveillance) are often prioritized. Meanwhile, counter-disinformation strategies often equate AIGC with traditional forms of disinformation, and it is often assumed that current tactics can be simply extended such that there are little to no explicit policies or strategies aimed directly at AIGC as a distinct problem requiring specific responses, as many experts have called for.

This ambiguity around the issue of disinformation further extends to how the problem is conceptualized more broadly. In terms of scale, disinformation can be understood as a problem in which harmful individual content spreads naturally between online users and thus requires more robust moderation mechanisms; such is the approach of the UK Online Safety Act. However, the national strategies of countries such as Spain (Gobierno de España, (2019) and France (Ajji, 2020) conceptualise disinformation as a coordinated and motivated campaign involving the spread of harmful narratives through numerous pieces of online content and thus require a national response. Furthermore, many of these strategies focus on the issue of electoral interference while overlooking the continual role that disinformation plays in everyday abuse, encouraging polarization between communities, and eroding confidence in democratic institutions.

With these different conceptualizations of disinformation comes further ambiguity around what constitutes harmful content. Notably, the UK Online Safety Act identifies harmful content as that which causes psychological or physical harm upon an individual, while the Digital Services Act (DSA) considers the broader societal harms of disinformation and other national criminal codes, such as those in Italy, Spain, and Albania, characterise harm in terms of public order and citizen safety.

Most critically, counter-disinformation policy must navigate the fundamental tension with freedom of speech. The boundary between harmful disinformation and protected speech is often blurred, and any policy, even one that is non-legislative, runs the risk of creating a chilling effect on legitimate expression. As discussed, a focus on banning or removing content can lead to further public distrust in regulatory institutions and can be easily co-opted by authoritarian regimes to suppress dissent.

Given these complexities and ambiguities, existing laws addressing harmful online content must be updated to address the specific challenges of harmful AIGC, and particularly, they require a clearer definition of what constitutes disinformation and what constitutes harm. We propose establishing a taxonomy of disinformation based on the semiotic models discussed in Chapter 3 and clearly identifying AIGC within this taxonomy. Such a taxonomy differentiates disinformation that is based on falsification of the material form (e.g., manipulation or fabrication) and that which is based on falsification of the content (e.g., misrepresenting authentic content). Harmful AIGC falls into the first category. Based on these categories, more specific definitions and guidelines can be established.

As the EPRS recommends, clearer guidelines are necessary for applying the General Data Protection Regulation (GDPR) framework to deepfakes, while strengthening the capacity of data protection authorities to address unlawful data processing, and developing a unified approach to personality rights within the EU (discussed below). Furthermore, we should protect the personal data of deceased persons, for example, with a "data codicil" and institutional support for victims of AIGC by providing accessible judicial and psychological resources.

Given the role that AIGC plays in individual harms (e.g., pornographic abuse), collective harms (e.g., political polarization), and societal harms (e.g., distrust in institutions), a unified strategy is crucial to addressing all three levels.

3.2 Unified Personality Rights

Similarly to definitions of AIGC and harms, personality rights covering an individual's name, likeness, image and voice are currently not harmonized at the EU level. This leaves regulation to the discretion of Member States and resulting in a patchwork of approaches. For example, France protects personality rights primarily through privacy and image rights, while Germany provides stronger safeguards by recognising personality rights under its constitution. By contrast, the UK lacks standalone legislation to cover personality rights but, instead, relies on a combination of privacy law, defamation, and tort law.

As the harms of AIGC transgress national boundaries, the EU should harmonize regulations related to personality rights to ensure consistent protection of citizens and to prevent malicious actors from exploiting these regulatory differences. A potential grounding for EU-level personality rights could be the recently proposed amendment to the Danish Copyright Act that is explicitly designed to address the issue of AIGC and digital imitations (Denmark, 2023).

This draft law treats identity as intellectual property and aims to give citizens copyright-style rights over their own likeness, voice, and physical features. Under the proposal, citizens can demand the removal of AIGC, representing themselves, made without consent, and seek compensation, even if no reputational damage is proven. Online platforms would be legally required to take down such content once notified or face sanctions, while carve-outs remain for free expression uses such as parody and satire. The law also offers specific protection to performing artists against unauthorized digital reproductions of their work. Broadly, this approach could be expanded across the EU to give citizens an explicit legal mechanism for controlling their own likeness and for combating individual harms of AIGC.

This could be achieved by updating existing legislation. Firstly, the EU Copyright Directive should be updated to give citizens the right to their 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 create robust regulation for preventing misrepresentation through AIGC.

3.3 Transparency Obligations

While the AI Act introduces transparency obligations to clearly label deepfakes circulating on online platforms, further transparency obligations should apply to AI moderation and deepfake detection systems used by these platforms. As discussed in 8.1, these technological containment measures risk being perceived by the public as authoritarian attempts at censorship that police the truth and insist upon a single arbiter. Without transparency, the use of AI systems to restrict the spread of harmful content may backfire causing further public distrust of governments and organizations. To combat this, we propose that platforms be required to disclose how their AI moderation and deepfake detection systems operate. This transparency would allow users to understand how content is moderated and flagged, while also providing a basis for holding platforms accountable for their decisions. Clear procedures for labelling deepfakes and a robust appeal mechanism must be established to ensure fair treatment and protect legitimate uses of GenAI.

3.4 Unified Infrastructural Investment

All of these strategies depend on strong government and private organizations, nationwide organizational networks, substantial funding, and the technical infrastructure needed for implementation. While robust policy frameworks may succeed in developed nations with sufficient capacity, they are often unworkable in regions with low digital literacy, limited access to technology, and weaker government systems. This digital divide is a major barrier to a unified European approach, and a major challenge to the integrity of our broader information environment and leaves us all more vulnerable to harmful AIGC. We must address this divide through international cooperation and investment programmes that build foundational digital infrastructure and establish comprehensive regulatory systems.

Without such efforts, proposed solutions risk deepening existing inequalities and failing to address the global scope of the threat. The EPRS (van Huijstee et al., 2021) highlights one response: authentication systems that enable users to verify content through digital watermarks or registered information provenance, extending also to court evidence. It further recommends coordinated investment in AI systems for detection and prevention, alongside diplomatic measures and international agreements to deter foreign state actors, reinforced where necessary by economic sanctions. To close capacity gaps in organizations and developing nations, the EPRS also calls for investment in knowledge and technology transfer, and for both public and private entities to conduct their own risk assessments. Primarily, this measure addresses broader societal harms of deepfakes and synthetic media by seeking to give all Member States and institutions sufficient tools to tackle disinformation across borders.

3.5 Multi-stakeholder Coordination

As discussed in Chapter 2, harmful AIGC can rapidly spread throughout online networks, and so it is necessary to establish early-warning systems that integrate technical and human intelligence. A primary obstacle to effective counter-disinformation strategies is institutional dysfunction (e.g., different standards and definitions for disinformation) and a lack of collaboration between key stakeholders across society, such as platforms, governments, research institutions, and media organizations. For example, governments may be hesitant to share sensitive data with private companies, while platforms may be unwilling to share proprietary data with public research institutions. Policy can attempt to bridge these gaps by establishing neutral, third-party convenors and by creating a clear set of shared ethical principles that all parties agree to uphold. This lack of collaboration and coordination is also evident between local, national, and European-level organizations, where differing policies, jurisdictions, and resources create inefficiencies. Some states have sought to tackle this issue directly. Notably, Spain's Protocol to Combat Disinformation (Gobierno de España, 2021) emphasizes inter-agency cooperation, while the UK has introduced regional cybersecurity hubs to coordinate responses, primarily to cyber threats and to disinformation instances (UK Government, 2022). However, many other states suffer from a lack of coordination. In particular, this dysfunction hinders efforts to rapidly address large-scale infodemic scenarios involving AIGC.

To address this dysfunction, key government institutions, social media platforms, fact-checking groups, and media organizations at the local, national, and international levels should establish a unified counter-disinformation network. Such a network would enable a real-time infodemic alert system whereby harmful AIGC identified by one organization can be immediately flagged for review by all partnering organizations, and the network as a whole can launch simultaneous public awareness campaigns to highlight the infodemic risk to citizens. Such a network approach would foster a transparent, agile verification process that allows multiple perspectives to contribute without resorting to heavy-handed state policing of the truth. Furthermore, the interconnected and multi-level nature of this approach would more effectively tackle infodemic events by enabling rapid verification and widespread public communication. This creates a network effect of protection, where the detection of a single piece of harmful content by one entity contributes to the resilience of the entire ecosystem, thus moving from a fragmented and reactive response to a more proactive and coordinated defence.

Key to this counter-disinformation network is increased investment in local journalism and media organizations that are trusted within their immediate communities. With increased funding, local media could provide reliable firsthand reporting that feeds into national and international levels, while also playing a direct role in public communication and serving as trusted intermediaries between the local community and the wider information ecosystem. Such investment would also be bolstered by greater coordination with online platforms to ensure citizens receive localized news. Furthermore, the use of local media organizations instead of government communication hubs ensures independence and avoids authoritarian tendencies.

While challenging to implement due to institutional dysfunction and lack of resources, this network approach is an effective way to address the multi-level, cross-border, and cross-platform nature of disinformation threats.

3.6 Media and AI literacy

For decades, there has been a strong emphasis on building public resilience to political manipulation through media literacy initiatives at both national and EU levels. Such efforts remain essential to democratic resilience in the age of synthetic media, empowering citizens to be active and critical participants in socio-political discourse.

However, current initiatives often lack a specific focus on GenAI, and so literacy programmes need to evolve to respond to our continually changing information environment. As the EPRS recommends, AI literacy should be integrated into formal educational curricula from a young age in order to teach students how to critically consume synthetic media and how to analyse its production, purpose, and potential harms (van Huijstee et al., 2021).

This includes teaching citizens how to identify AIGC (e.g., unnatural eye movement, distorted backgrounds, audio glitches), as well as a broader understanding of how GenAI systems are trained and the biases they may contain. Moreover, literacy programmes should teach citizens to recognise AI-generated content based on technical and, furthermore, encourage citizens to consider the context, such as the content's source and broader background information about the people and events they are shown.

This does not simply require more general media literacy training, and requires citizens to be more deeply engaged with politics and events. Furthermore, AI literacy initiatives should engage citizens across all stages of life, from primary education to professional training and adult programmes. Meanwhile, targeted programmes should seek to engage vulnerable groups who may lack certain literacy skills, such as older adults or people with learning and cognitive disabilities.

Promoting AI literacy is not only an effective strategy for combating individual manipulation or deception, but, if implemented consistently across society, such initiatives address those broader epistemic and societal harms caused by disinformation. By equipping citizens with the ability to discern reliable information from synthetic noise, we can begin to rebuild trust in democratic institutions and political processes. While such initiatives should receive government funding, independent educational institutions and citizen science organizations must implement AI literacy programmes to avoid the perception of authoritarian arbitration of truth that Farkas and Schou highlight. Such programmes can lead to an AI-literate citizenry that is more resistant to manipulation. If coupled with technical citizenship initiatives, as the next section will explain, this could further encourage a more vibrant AI-enabled public discourse and political participation.

3.7 Technical citizenship

To encourage a more vibrant and active political participation, AI literacy programmes need to go beyond simply teaching ways of identifying AIGC and critical engagement with GenAI. These programmes should also focus on ethical and pro-democratic use of such technologies that do not focus on deceptive practices but, rather, methods of AI-enabled personal representation and self-expression. Investing in this more practical curriculum is to cultivate a citizenry that is AI literate and aware of the technology's societal impacts, and is also utilising AI positively and actively engaging in plural democratic debate. It is important for these initiatives not to simply encourage greater use of GenAI but to emphasise the ethical use of these technologies for personal representation and self-expression rather than manipulative deception.

Beyond further investment in formal education programs for technical citizenship, policy can be used to promote informal and community-driven initiatives. Policy support could include publicly funded online spaces or channels for teaching AI literacy and ethical use, as well as grants for community-based organizations to host workshops and information sessions, particularly in marginalized communities disproportionately affected by disinformation campaigns (Gautam et al., 2024). Such sessions could focus on creating online spaces wherein citizens can participate in political discussions in creative and empathetic ways by utilising AI-generated content. Platforms such as YouTube and GitHub could also be repurposed as such spaces for public engagement (McCosker, 2024).

Combined with media and AI literacy, technical citizenship initiatives are intended to encourage a more trustworthy information environment and to promote a pluralist media landscape in which citizens are politically engaged and where numerous different socio-political views are represented.

3.8 Pluralist media landscape

Beyond literacy and technical citizenship initiatives, a significant obstacle to implementing counter-disinformation strategies is the increasingly fragmented media landscape across Europe and within individual Member States. The widespread availability of digital technologies and the rapid growth of social media have drastically increased the number of people capable of producing and disseminating information online. As such, many users and entire communities no longer share common sources of information, instead consuming highly personalized content shaped by recommendation algorithms. This explosion of online platforms makes it difficult to monitor information flows and ensure compliance with counter-disinformation legislation. Notably, the provisions of the DSA only apply to very large online platforms, leaving smaller but still influential sources largely unregulated.

Countering disinformation requires a strong, diverse media ecosystem. Policymakers should support independent journalism and media organizations to ensure that the public has access to reliable, high-quality information, while also supporting pluralistic debate. Promoting diverse media sources and critical reporting can help resist the normalization of biased or distorted narratives through AIGC, without resorting to authoritarian overreach.

A key component of this approach is addressing capacity gaps that exist in smaller media organizations and civil society groups that are essential for ensuring diverse perspectives. Policy could establish national or international funds, supported by government grants and philanthropic contributions, to provide these organizations with access to advanced tools and training. This would ensure that the ability to combat disinformation is not a luxury reserved for well-funded entities, but a widely distributed capability that strengthens the entire information ecosystem. Crucially, this approach avoids the centralization of media power, instead fostering a plural and resilient information ecosystem.

4 Concluding Remarks

Any comprehensive strategy that aims to effectively regulate against the harms of AIGC in the European context must first recognise that these harms are rooted in the degradation of our information environment. Accordingly, the harms posed by AIGC are not solely related to misrepresentation or deception of individuals, but rather they relate more broadly to the integrity of collective knowledge and manifest differently across different levels of society (individual, collective, societal).

Existing EU legislation remains fragmented and inadequate when addressing this specific issue, and there is an urgent need for more clarity. However, legal tools alone are insufficient to address the deep social integration of these technologies into our social lives and the diverse harms this integration presents. Additionally, these legalistic approaches do not fully embrace the potential opportunities for using GenAI to revitalise plural political debate. To properly address this issue, policymakers should adopt a holistic approach that balances technical and legal solutions aimed at containing disinformation with pluralist social policies aimed at promoting political participation.

In this chapter, we developed an approach oriented around three primary strategic objectives: (i) clarifying harms of AI-generated content through unified legal definitions and personality rights; (ii) strengthening institutional coordination through multi-stakeholder collaboration and investment; and (iii) enhancing citizenship through AI literacy, technical skills, and a plural media landscape. Rather than viewing AIGC solely as a threat to be contained through heavy-handed measures, regulatory and policy innovations should focus on adapting society around GenAI. Central to future democratic resilience is the cultivation of a technically literate and politically active citizenry that is able to recognise and resist AI-generated disinformation and actively uses GenAI tools to contribute to the political debate.

End notes

Andrew McIntyre conceptualized the Chapter and coordinated the writing. He wrote the introduction and conclusions, deepfakes for political participation, Regulatory and policy priorities for democratic resilience. Yasaman Yousefi wrote the section on Policy and Pluralism. She also contributed to the section on Regulatory and policy priorities for democratic resilience, specifically to the legal analysis

and policy recommendations. Maria Dolores Sánchez Galera contributed to the legal analysis. All authors reviewed and approved the final version.

References

Ajji, K. (2020). Protecting liberal democracy from artificial information: The French proposal. In B. Petkova & T. Ojanen (Eds.), Fundamental rights protection online (pp. 57–83). Edward Elgar Publishing. https://doi.org/10.4337/9781788976688.00013

Allen, D., & Weyl, E. G. (2024). The real dangers of generative AI. *Journal of Democracy, 35*(1), 147–162.

Bisconti, P., McIntyre, A., & Russo, F. (2024). Synthetic socio-technical systems: Poiêsis as meaning making. *Philosophy & Technology, 37*(3), 94. https://doi.org/10.1007/s13347-024-00730-y *(DOI added when available – remove if you prefer strictly source-based)*

Codice Penale (Italia), R.D. 19 ottobre 1930, n. 1398, arts. 656, 595, 612.

Criminal Code of the Republic of Albania, Law No. 7895 of 27 January 1995, as amended by Law No. 146/2020, arts. 267, 271.

Denmark. (2023). Consolidated Act on Copyright (Consolidated Act No. 1093 of August 20, 2023). WIPO Lex. https://www.wipo.int/wipolex/en/legislation/details/22692

Doherty, B. (2023, April 8). Manus Island and Nauru: Previously unseen testimony and AI imagery reveal "unimaginable" part of Australian history. *The Guardian*. 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

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Official Journal of the European Union, L 168*, 1–135.

Farkas, J., & Schou, J. (2020). *Post-truth, fake news and democracy: Mapping the politics of falsehood.* Routledge.

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

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality.* Oxford University Press.

Gautam, A., Joshi, R. K., Narula, A., & Sharma, N. (2024). Mitigating human rights violations caused by deepfake technology. *Library Progress (International), 44*(3), 4628–4637.

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

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

Jee, C. (2020, February 19). An Indian politician is using deepfake technology to win new voters. *MIT Technology Review*. https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/

Laclau, E. (1990). *New reflections on the revolution of our time.* Verso.

Lee, D. (2019, May 10). Deepfake Salvador Dalí takes selfies with museum visitors. *The Verge*. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum

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

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

McCosker, A. (2022). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *New Media & Society, 26*(5), 2786–2803.

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)*

Mouffe, C. (1993). *The return of the political.* Verso.

Nemitz, P. (2022). People or technology: What drives democracy? *Transatlantic Policy Quarterly, 20*(4), 35–42.

Rancière, J. (2014). *Hatred of democracy*. Verso.

Risse, M. (2023). *Political theory of the digital age: Where artificial intelligence might take us*. Cambridge University Press.

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

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

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

UK Government. (2022). *Government Cyber Security Strategy: Building a cyber resilient public sector.*

UK Government. (2023). *Online Safety Act.*

Van Huijstee, M., van Boheemen, P., & Das, D., et al. (2021). *Tackling deepfakes in European policy*. European Parliamentary Research Service.

University of Maribor Press

Conclusion: Whence and Whither of AI-Generated Content

YASAMAN YOUSEFI,[1, 2] FEDERICA RUSSO[3]
[1] DEXAI-Artificial Ethics, Rome, Italy
yasaman.youefi@dexai.eu
[2] University of Bologna, CIRSFID ALMA AI, Faculty of Legal Studies, Bologna, Italy
y.yousefi@unibo.it
[3] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
f.russo@uu.nl

AI-generated content represents a multifaceted and complex contemporary issue. Nowadays, technologies to generate text and audiovisual contents are at the fingertip of the general public.

The emergence of synthetic media and the widespread use of deepfakes in multiple contexts, alongside the decline of reliance on traditional journalism as a news source, has been shown to undermine democratic processes, from geopolitical stability to interpersonal trust. The SOLARIS project has studied the emergence of these technologies and phenomena, provided a theoretical lens through which analyse their societal relevance and impact, run empirical studies and activities with multiple stakeholders, and finally formulated policy recommendations to address the challenges that synthetic media pose. In this closing chapter, we briefly recall the results of our investigations.

As is recalled in Chapter 1, Generative AI models (such as GANs or Diffusion Models) have a long tradition in computer science and in Artificial Intelligence and are the result of decades of investment in research and development. In the past couple of years, however, there has been a fundamental change: the quality of the produced outputs has dramatically improved, *and* the technologies have become more widely available and easier to use, also without a technical background. The so-called "deepfakes" thus represent a paradigm shift in the disinformation phenomenon because they reduce the costs of creating highly persuasive, realistic data to near zero, profoundly altering the information supply chain.

Changes in the production and circulation of AI-generated content through online networks are documented in Chapter 2. Threats to the information ecosystem are not episodic anymore, but are becoming *systemic*, driven by the interaction between human psychology (exploiting emotional responses), platform incentives (algorithmic amplification), and organized network activity (influencers and botnets). The findings from SOLARIS Use Case 2, presented in Chapter 2, confirm that human expertise and contextual knowledge remain non-negotiable safeguards, requiring an integrated strategy of technology, regulation, and education to succeed. Safeguarding democratic discourse requires defeating the initial viral spread, which demands platform accountability and sophisticated early-warning systems that integrate both technical and human intelligence. And yet, these forms of de-bunking and pre-bunking are on their own insufficient.

We explain why that is the case in Chapter 3, in which we reconceptualise deepfakes, and more generally AI-generated content, as part of complex socio-technical systems. Through the lenses of Actor-Network Theory, we laid down the "hybrid" network that makes the generation, circulation, and spreading of such synthetic media possible. It is a hybrid network that involves the Tech Industry, social media platforms, the public, and society, both as users and recipients and as potential targets, institutions, and legislative mechanisms. Focusing on the "artefact" only is insufficient to explain why these synthetic media are profoundly changing modes of communication. From a *semiotic* perspective, in particular, we can appreciate why this novel way of altering pictures and videos poses challenges to democratic processes that older software such as Photoshop did not. Briefly and simply put, modifications are not just aesthetic but involve intentions to deceive and produce narratives able to influence public discourse in unprecedented ways. This is because of the speed at which such content can circulate and, foremost, because of how we interact with these media.

In Chapter 4, we present the findings of our empirical study (Use Case 1) on the "psychology of deception". We developed a psychometric scale of "perceived trustworthiness" that allows us to measure attitudes of users towards which shows why even poor-quality deepfakes can be considered highly realistic. The methodology and the findings are thoroughly presented in that chapter and in dedicated scientific publications. Briefly put, the *quality* of the produced media is not the only factor lending to credibility. The message, the contexts, our prior beliefs related to the subject, as well as our own individual media environment, are all elements that (partly) explain why we come to believe in deepfakes.

The consequences for democracy are not difficult to foresee. In Chapter 5, we explain how deepfakes can easily become "political weapons." But the word "political" here has a larger meaning than just politics *strictu sensu*. 2024 has been a crucial year globally, in which we have witnessed how much political campaigns have made use of deepfakes in a very subtle way: public opinion and evidentiary truth are easily manipulated. But equally important, the deployment of deepfakes, particularly through targeted campaigns against women and minority candidates, is shown to exacerbate structural inequalities, indicating that the technology acts as a force multiplier for existing societal biases.

The narrative that deepfakes are a threat to our democratic society, and fuelling the modes of working of totalitarian ones, contributing to information warfare, has quickly gained traction. These threats are real, but SOLARIS has made an effort *not* to demonise these technologies and to explore their potential for *good* use. In Chapter 6, we report on the activities of Use Case 3. We produced AI-generated videos to support and disseminate selected Sustainable Development Goals. In line with the semiotic approach developed in Chapter 3, we discussed with citizens the features that make such synthetic media more credible and effective. Having a "good" message is not enough, and in fact, as we also discuss in the chapter, it is ethically controversial and contentious to establish what "good" is. But the significant result of our activities is that, once again, (technical) quality of audiovisual contents is insufficient, on its own, to successfully deliver a "good" message. We need instead to give attention to how the narrative is built, which means engaging with several parameters (first vs third person, a real person vs an AI-generated avatar, etc.). In short, AI for good *is* possible, but it requires way more than just technical abilities.

The complex and nuanced landscape that SOLARIS depicted is mirrored in the equally complex area of legal initiatives and regulatory efforts. In Chapter 7, we discuss how to govern deepfakes and synthetic media. We have found that governing deepfakes requires more than just tech fixes or legal changes. It needs instead an integrated approach, in line with the network approach sketched in Chapter 3. Successful regulation hinges on being extremely clear about the harms (conceptual clarity) and embracing a principle of relational responsibility to tackle the immaterial consequences of AI-generated Content, which thrive in the current sociological atmosphere of "existential relativism," where truth feels entirely blurry. Future policy must be proactive, focusing on due diligence obligations for platforms and holding actors accountable for harms to the information ecosystem itself, rather than solely on individual cases of deception.

In Chapter 8, we further develop on what, according to SOLARIS findings, is the way to go. We argue that the most effective strategy for democratic resilience is a shift from reactive moderation and containment to proactive empowerment and pluralism. The three-pronged approach clarifying AI-generated content's harms and rights, strengthening institutional coordination, and enhancing citizenship, reconceptualizes Generative AI not as a threat to democracy, but as a potential resource for fostering more inclusive political dialogue and a trustworthy information environment. Policy should champion citizens' ability to critically engage with and create synthetic media, demanding a regulatory focus on rights clarification (especially personality rights) and mandatory AI literacy standards.

In sum, understanding synthetic media and the technologies able to generate them requires a cross-disciplinary, socio-technical multi-stakeholder approach, such as the one pursued by SOLARIS, grounded in the Actor-Network Theory framework. This approach is able to take into account *all* the actors involved, the design of AI technologies, and the level of literacy and awareness in the general population, as well as ethical, legal, and policy considerations.

The real challenge of deepfakes to modern democracies is how *δῆμος*[1] can be put *again* at the centre. We should therefore not just consider society as an "undifferentiated" whole; societies are made of citizens. Individual *and* collective levels are needed to understand potential harms and benefits, as well as routes for regulation and use. A harm-based approach, such as the one advocated in the AI Act, represents an initial, valuable guidance for the ethically oriented evaluation of the normative implications of synthetic media. But more is needed. As previously discussed, AI content generation should not be demonized. Instead, the challenge lies in the identification of the correct ways of ensuring fairness and well-being in desirable AI-generated content that is used for civic engagement and participation-potentially an AI for good *and* to simultaneously identify ways to limit the generation and spreading of potentially harmful content.

Our proposed policy approach, therefore, builds on the core insight that democratic resilience in the age of generative AI cannot be achieved through containment alone. Instead, it requires a reorientation of governance from reactive truth-policing to proactive democratic cultivation. We articulated this shift in Chapter 8 through a three-pronged framework: clarification, coordination, and citizenship, which together operationalise a harm-based and network-oriented understanding of AI-generated content.

Clarification responds to individual and societal harms by establishing legal certainty around synthetic media, liability, personality rights, and transparency obligations, thereby restoring agency and accountability in increasingly ambiguous communicative environments. Coordination addresses systemic and societal risks by strengthening institutional infrastructures, fostering cross-sector collaboration, and reducing epistemic asymmetries across Member States, enabling proportionate and timely responses to high-impact manipulations. Citizenship, finally, targets individual and collective harms by investing in AI and media literacy, technical citizenship, and pluralist media ecosystems, recognising citizens not merely as potential victims of deception but as active interpreters, co-creators, and ethical agents within the infosphere.

[1] From the Greek word dêmos, the word means 'the people', as in compounds like dēmo-kratia, "people-power" or "democracy".

Taken together, these principles reflect a normative commitment to "more politics," rather than less: a vision of democratic governance that accepts contestation, plurality, and uncertainty as constitutive features of the public sphere, rather than pathologies to be eliminated. By embedding generative AI within participatory, transparent, and pluralistic structures, this approach seeks to protect democratic institutions while simultaneously expanding civic capacity. The aim is to ensure that technological innovation reinforces collective self-determination rather than displacing it. In this sense, SOLARIS generative AI as a site of democratic struggle and possibility, one in which resilience emerges from empowered citizens, coordinated institutions, and clearly articulated rights and responsibilities.

Endnotes

Yasaman Yousefi and Federica Russo equally contributed to the writing of the conclusions.

Deepfakes, Democracy, and the Ethics of Synthetic Media: A Synthesis of the SOLARIS Project

Yasaman Yousefi[1, 2] Lucy Conover,[3] Izidor Mlakar,[4] Federica Russo[5] (eds.)

[1] DEXAI-Artificial Ethics, Rome, Italy
yasaman.youefi@dexai.eu
[2] University of Bologna, CIRSFID ALMA AI, Faculty of Legal Studies, Bologna, Italy
y.yousefi@unibo.it
[3] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
l.a.conover@uu.nl
[4] University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
izidor.mlakar@um.si
[5] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
f.russo@uu.nl

This book presents an interdisciplinary synthesis of research findings from the Horizon Europe project SOLARIS. It systematically examines deepfake technologies as a structural threat to democratic institutions. Eight chapters integrate technological, psychometric, semiotic, legal, and political science perspectives, empirically analyzing not solely epistemic risks, but also potential positive applications of synthetic media. The authors critically examine generative neural network architectures, viral propagation mechanisms of disinformation in digital networks, and cognitive-affective factors in perceived trustworthiness of AI-generated content. The work identifies deepfake impacts on electoral integrity, epistemic erosion of public discourse, and asymmetric effects on marginalized populations. Critical evaluation of European regulatory instruments (AI Act, DSA) leads to formulation of innovative policy recommendations for systemic resilience of democratic processes. The research also documents positive implementations of synthetic media in pedagogical, cultural-documentary, and civic engagement contexts. This publication mainly targets the research community, policymakers, media professionals, and technological actors.

DOI
https://doi.org/
10.18690/um.feri.2.2026

ISBN
978-961-299-109-8

Keywords:
synthetic content,
democracy,
generative artificial intelligence,
deepfakes,
disinformation,
epistemic erosion,
ethics in artificial intelligence,
media literacy,
democratic resilience,
synthetic media for social good

University of Maribor Press

In the digital age, the proliferation of disinformation and misinformation has been exponentially amplified by tools that enable rapid dissemination. The media landscape has undergone a paradigm shift with the emergence of artificial intelligence technologies capable of generating synthetic images, videos, and speech, collectively known as deepfakes. These sophisticated digital fabrications challenge traditional paradigms of trust, truth, and transparency, with profound implications for democratic governance and societal cohesion. *Deepfakes, Democracy, and the Ethics of Synthetic Media: A Synthesis of the SOLARIS Project* provides a comprehensive examination of this phenomenon, offering a multidisciplinary analysis of the societal, ethical, and political challenges posed by deepfakes while exploring potential pathways to harness their benefits responsibly.

This volume integrates contributions from the EU-funded SOLARIS project and related initiatives, synthesizing expertise from academia, policy, and technology. It aims to empower policymakers, technologists, educators, and the public with actionable insights and informed approaches to mitigate risks while leveraging the transformative potential of synthetic media.

SOLARIS

Funded by
the European Union

For more information,
see SOLARIS website

Univerza v Mariboru

Faculty of Electrical Engineering
and Computer Science