SOLARIS

Strengthening democratic engagement through value-based generative adversarial networks

D6.1 SOLARIS
Differentiating GANs impact and implication on different social groups and diversifying policy options

Lead Authors: Calogero Caltagirone, Angelo Tumminelli (LUMSA University)

With contributions from:
AMI CINI, DEX, ECSA UC3M, UM, UU
Reviewers: ANSA, UvA

Deliverable nature:	R-Document
Dissemination level: (Confidentiality)	PU
Delivery date:	September 2025
Version:	Final
Total number of pages:	57
Keywords:	Digital Citizenship, Digital Education, ANT (Actor-network theory), Inclusion, Digital Divide, Human Fulfilment

Executive Summary

This deliverable (D6.1) explores how Generative AI (GenAI) and synthetically generated social actors can contribute to strengthening digital citizenship and democracy. Beginning with education, the report highlights the transformative potential of GenAI in personalizing learning, fostering creativity, and promoting inclusion at all levels of schooling. At the same time, it warns that such benefits come with risks of bias, manipulation, and ethical misuse, which make digital literacy, civic skills, and critical thinking essential for preparing future citizens.

The analysis then turns to the societal and legal implications of GenAI, particularly deepfakes. These technologies threaten trust, accountability, and human rights, disproportionately harming vulnerable groups such as women, ethnic minorities, youth, and the elderly. Despite the urgency of these risks, current legal frameworks remain underdeveloped and fragmented, with little empirical or interdisciplinary research to guide accountability and regulation. Actor–Network Theory (ANT) is introduced as a lens for understanding GenAI not only as a technical tool but as a social actor that shapes meaning, values, and power relations within democratic life.

Building on this perspective, the deliverable examines the ethical use of synthetically generated social actors. Deepfakes of real individuals, whether living or deceased, raise acute questions of dignity, autonomy, and privacy, while non-existent AI avatars offer less harmful alternatives if transparency and accountability are ensured. These reflections connect directly with the project's co-creation work: in Use Case 3, citizens actively participated in producing value-driven GenAI content. This participatory approach was shown to foster transparency, strengthen digital literacy, and build trust, demonstrating that democratic engagement is best achieved when citizens themselves help shape the technologies that affect them.

Finally, the report applies semiotic analysis to develop a taxonomy of synthetic media, clarifying how GenAI content conveys meaning, credibility, and authority. This provides a framework for both educational strategies and regulatory design. The overall findings are clear: GenAI can serve as a powerful tool for education, inclusion, and citizen engagement, but it also poses profound risks to democracy when used irresponsibly. Vulnerable groups are particularly exposed to harm, while the lack of strong interdisciplinary research and legal accountability frameworks leaves significant gaps in protection.

In response, the deliverable recommends clear labelling and transparency obligations for AI-generated content, investments in digital literacy programs, participatory co-design processes to ensure societal alignment, and the development of accountability frameworks that combine legal, technical, and social expertise. Taken together, these measures underline the central conclusion: only a multi-faceted, participatory, and ethically grounded approach can ensure that GenAI contributes to democratic engagement while mitigating its risks and safeguarding human dignity.

Document information

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

Deliverable	**Number**	6.1	**Title**	Differentiating GenAI impact and implication on different social groups and diversifying policy options
Work package	**Number**	6	**Title**	Enhance capacities for digital citizenship and democracy
Task	**Number**	T6.1, T6.2	**Title**	Methodologies for enhancing digital citizenship with synthetically generated social actors, segmenting GANs impact on social media Across different socio-economic groups from the legal and psychological point of view

Date of delivery	**Contractual**	M32	**Actual**	M32
Status	version 1		☒Final version	
Nature	☒Report ☐Demonstrator ☐Other ☐ORDP (Open Research Data Pilot)			
Dissemination level	☒Public ☐Confidential			

Authors (partners)	Andrew McIntyre (UvA) Angelo Tumminelli (LUMSA) Antonio Estella (UC3M) Calogero Caltagirone (LUMSA) Domenico Bloisi (CINI) Giuditta Bassano (LUMSA) Izidor Mlakar (UM) Lek Hakani (AMI) Letizia Aquilino (DEX) Maria Dolores Sanchez Galera (UC3M) Michele Brienza (CINI) Nefertari Nachtigall (ECSA) Nejc Plohl (UM) Piero Polidoro (LUMSA) Romane Blassel (ECSA) Tommaso Tonello (UU) Yasaman Yousefi (DEX)			
Responsible author	**Name**	Angelo Tumminelli, Calogero Caltagirone		
	Partner	Lumsa University	**E-mail**	a.tumminelli@lumsa.it c.caltagirone@lumsa.it

Summary (for dissemination)	In this deliverable, we report on findings issued from WP6 activities. SOLARIS finds that GenAI can enhance digital citizenship and democratic engagement but also poses risks, especially for vulnerable groups such as women, minorities, youth, and the elderly. Using segmentation strategies,

	Actor–Network Theory, and a co-creation methodology for Use Case 2, we highlight both the societal and legal challenges of deepfakes and propose ethical best practices and policy options. Overall, we conclude that citizen participation, digital literacy, and transparency are key to maximising GenAI benefits while limiting potential harms.
Keywords	Digital Citizenship, Digital Education, ANT (Actor-network theory), Inclusion, Digital Divide, Human Fulfilment

Acknowledgement

Disclaimer

Table of contents

Abbreviations and acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
AI4SG	Artificial Intelligence for social good
ANT	Actor–network theory
AIGC	Artificial Intelligence Generated Content
CA	Consortium Agreement
CC BY 4.0	Creative Commons Attribution 4.0 International
Celeb-DF, DFD, DFDC	Publicly available datasets developed for deepfake detection
CERN	European Organisation for Nuclear Research
DDI	Data Documentation Initiative
DMP	Data Management Plan
DOI	Digital Object Identifier
DPO	Data Protection Officer
EC	European Commission
EU	European Union
FAIR data	Data which meet principles of findability, accessibility, interoperability, and reusability
GANs	Generative Adversarial Networks
GenAI	Generative Artificial Intelligence
GDPR	General Data Protection Regulation
UC	Use Case

1. Role of GenAI in Education and Digital Humanism

The rapid diffusion of generative artificial intelligence (GenAI)1 has brought new urgency to long-standing debates about education, ethics, and citizenship in the digital age. Education no longer concerns only the transmission of knowledge and skills but also the cultivation of digital literacy, ethical responsibility, and civic participation within the broader "infosphere" (Floridi, 2004; 2016). In this environment, GenAI reshapes both opportunities and risks: it can enhance learning through personalization, creativity, and inclusion, yet it also raises critical challenges of bias, manipulation, and democratic erosion. Anchoring these dynamics within the tradition of digital humanism, this section examines how the exercise of citizenship is redefined in a context where rights, duties, and freedoms are increasingly mediated by algorithmic systems. It outlines the ethical values that sustain responsible digital citizenship, while exploring how education can foster the skills and dispositions needed to navigate, critically engage with, and ethically shape the digital infosphere.

1.1 Digital citizenship in the context of Infosphere

This section of D6.1 is devoted to defining digital citizenship in the context of the infosphere and identifying the main ethical values associated with the exercise of freedom using GenAI. It should first be pointed out that a citizen is a member of a political community, and he/she enjoys the rights and assumes the duties of membership. This broad definition is discernible, with minor variations, in the works of contemporary authors (cf. La Torre, 2022; Balibar, 2012; Leydet, 2010; Kymlicka & Norman, 2000) as well as in the entry "citoyen" in Diderot's and d'Alembert's *Encyclopédie* (1753). Notwithstanding this common starting point and certain shared references, the differences between 18th century discussions and contemporary debates are significant. The encyclopédiste's main preoccupation, understandable for one living in a monarchy, was the relationship between the concepts 'citizen' and 'subject'. Were they the same (as Hobbes asserted) or contradictory (as a reading of Aristotle suggested)? (cf. Hobbes, 1991, 148) This issue is less central today as it is often assumed that a liberal democratic regime is an appropriate starting point for thinking about citizenship (Costa, 2005; Hutchings & Danreuther, 1999, Kymlicka, 1995; Taylor, Turner, & Hamilton, 1994; Habermas, 1991).

In particular, the analysis here focuses on the exercise of citizenship in the digital age characterised by the pervasiveness of AI in people's lives. It should first be made clear that *Digital citizenship* refers to the set of rights, duties, skills, and behaviours that every person must adopt to actively, consciously, and responsibly participate in digital society (Ceccarini, 2021; Isin & Ruppert, 2015; Mossberger, Tolbert, & McNeal, 2007). In other words, it means knowing how to use digital technologies (the internet, social media, online services, etc.) in a safe, ethical, and constructive way (Dahal, 2023; DeHart, 2023). In this sense, the key aspects of digital citizenship are:

1. *Digital access and inclusion*: ensuring that everyone has access to technology and information.

[1] In earlier phases of research, **Generative Adversarial Networks (GANs)** were the dominant reference point for synthetic media. However, the technological landscape has rapidly evolved, and today a broader set of **Generative AI (GenAI)** models—including diffusion models, transformers, and large multimodal systems—hold a key role in creating synthetic content. To reflect this shift, this deliverable adopts the term *GenAI* rather than limiting the discussion to GANs. This choice ensures that the analysis captures the wider family of generative technologies shaping education, democracy, and digital citizenship, while remaining aligned with current academic and policy discourse.

2. *Digital literacy*: having the skills to use digital tools effectively.
3. *Online safety*: protecting personal data, privacy, and recognizing cyber threats.
4. *Ethical behaviour*: respecting others online, avoiding hate speech, cyberbullying, and fake news.
5. *Active participation*: using digital tools to contribute to society (e.g., signing petitions, engaging in online discussions, etc.).
6. *Legal responsibility*: understanding the laws and regulations governing digital use.

A good digital citizen knows how to navigate the web safely, contributes positively to the online community, and respects the rules and values of civil coexistence, even in the virtual world. The philosopher Luciano Floridi is one of the leading scholars in the philosophy of information and digital ethics. In the context of digital citizenship, Floridi addresses topics such as the infosphere, digital responsibility, and rights in the information age. Floridi defines the infosphere (cf. Floridi, 2004, 2016) as the global information environment where humans and digital technologies interact. Being digital citizens means consciously inhabiting this infosphere and contributing to its quality. Digital citizenship requires moral responsibility: it is not enough to simply follow laws; individuals must act ethically online, avoiding misinformation, cyberbullying, and privacy violations. Floridi argues that in the digital era, we need new rights to protect individuals, such as the right to be forgotten, transparency, and data protection. Algorithms influence our online experiences (from social media to Google searches). Floridi warns that an aware digital citizen must understand and question the power of algorithms to avoid manipulation and informational bias. According to Floridi, digital citizenship is not just a right but also a skill to be developed through digital literacy. People must learn to manage their online identity, recognize fake news, and critically use technologies. In summary, for Floridi, digital citizenship is not just about accessing technology but about a way of being and acting in the digital world with awareness, ethics, and responsibility (Pascucci, 2021; Rodotà, 2014; Pollicino, Bertolini, & Lubello, 2013; Ziccardi, 2012; Cogo, 2010).

Luciano Floridi also explicates the applicative potential of the political use of AI for Social Good (AI4SG), pointing out that ethical use of AI necessarily implies 'the design, development and implementation of AI systems in such a way as to (I) prevent, mitigate or solve problems that negatively affect human life or/and the wellbeing of the natural world and/or (II) enable socially preferable and/or environmentally sustainable developments'. Floridi underpins his reflection on AI ethics with five principles, borrowed from an established ethical framework in bioethics, with the aim of combining the use of AI and the promotion of both the individual and the common good of humanity. According to Floridi, the principle of beneficence requires the creation of an AI technology that is beneficial to humanity and that focuses on promoting the wellbeing of people and the planet, thus preserving human dignity in 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 so as to avoid the misuse of AI technologies that could harm humanity as a whole. The principle of autonomy, then, is the one that is called upon to safeguard the freedom of individuals as a shared patrimony: if it is true that when adopting AI and its intelligent action, the individual voluntarily cedes part of his or her decision-making power to machines, affirming the principle of autonomy in the context of AI means achieving a balance between the decision-making power that the individual retains within himself or herself and that which he or she delegates to artificial agents. From this, not only should human freedom be promoted, but also machine autonomy should be restricted and made

inherently reversible. Then there is the principle of Justice and the principle of Explicability, which have even more obvious repercussions in the social and global sphere: the former refers to the possibility of using AI to eliminate social discrimination by ensuring the consolidation of existing social structures and an adequate distribution of strategic resources; the latter concerns, finally, the need for AI to be able to 'make itself understood' as to what good or harm it is actually doing in society and in what ways; for AI to be able to promote and not limit human freedom and the common good, its use must be informed and aware, i.e. based on an awareness of its scope on both a personal and societal level.

Floridi's perspective is of particular interest because it places at the centre of an ethical use of AI the social good and the possibility of its realisation through personal freedom. In this sense, both Pope Francis's admonition and Floridi's criteria pose a well-defined toolkit for an AI use that can promote the human being by putting him in a position to express himself in the fullness of his value. And it is only when this happens in a society that the common good and peace are realised: this is not a utopia but an ethical task that awaits all human beings in the face of the challenges of their time (cf. Floridi, 2002, 227).

It should also be stressed that the influence of AI on social good, citizenship, and democracy is directly proportional to the protection/violation of certain human rights. Freedom of thought is one of the main rights of a democracy as well as the fundamental tool for building peace within just societies: people must be able to think freely without being subjected to more or less obvious forms of ideological or identity manipulation. The social good, in fact, is founded on pluralism and the difference of viewpoints that are the indispensable pillars of a democratic society. AI systems have the power to stimulate man's creative thoughts, presenting concepts that some might not have considered, but, at the same time, they risk ideologically orienting personal identities by conforming them to homologated standards of thought devoid of critical reflection.

As is well known, GenAI, with its power to generate deepfake images and videos, can put people in situations they have never been in, manipulate the perception of events, and potentially, manipulate an entire society, even by altering the processes of self-identification and socio-political belonging. It then becomes crucial to be able to discern the possibilities and limits of the use of AI for social freedom (Cassano & Contaldo, 2009). Dynamics such as information manipulation and ideological propaganda, mass surveillance and restriction of privacy, electoral interference and undermining of democratic decision-making, together with censorship and control of access to information, are just some of the ethically problematic aspects related to the social use of AI as a widespread tool (Cf. Amoroso & Tamburrini, 2018)

In short, there is a close connection between AI and citizenship: it is not possible today to think about the exercise of freedom outside the context of the infosphere.

It is necessary to ask, at this point, what ethical values are associated with the exercise of digital citizenship. We propose that the ethical values of digital citizenship are principles and behaviours that guide the conscious, responsible, and respectful use of digital technologies. These include:

1. **Respect***:* Interacting online with respect for other users and their rights. This means avoiding aggressive behaviour, insults, cyberbullying, or discrimination, and promoting an inclusive and tolerant communication environment.
2. **Responsibility**: Acting with awareness, considering the consequences of one's online actions. This includes being accountable for the content one shares, protecting personal information, and respecting the privacy of others.
3. **Safety**: Taking measures to protect one's own safety and that of others online. This involves using secure passwords, keeping devices updated, preventing fraud and scams, and raising awareness about digital risks.
4. **Honesty and Integrity**: Avoiding the spread of false or misleading information (such as fake news) and respecting copyright. This also means citing sources, giving credit to original authors, and avoiding plagiarism.
5. **Inclusion**: Promoting fair access to digital resources and encouraging participation from everyone, overcoming digital barriers like discrimination and the digital divide.
6. *Empathy and Compassion*: Understanding the experiences and feelings of other users, showing kindness and understanding in online interactions.
7. **Privacy Awareness**: Respecting one's own privacy and that of others, handling personal information carefully, and ensuring it is not misused or shared without consent.
8. **Digital Sustainability**: Sustainably using technology, minimizing the environmental impact associated with excessive or irresponsible use of technological resources.
9. **Active Citizenship**: Actively participating in the digital community to improve and promote the common good, for example, by contributing educational or informative content, supporting social causes, and respecting digital regulations (Cf. Aneesh, A., 2006).

These ethical values support responsible digital citizenship, essential for building a safe, respectful, and inclusive digital community.

Digital citizenship is the competence and commitment to participate responsibly, consciously, and actively in digital life and online interactions. It involves the appropriate and responsible use of digital technologies, considering rights, duties, and ethical standards to ensure a positive and respectful online presence. Being a "digital citizen" means understanding and applying the rules of online communication, respecting others' privacy and security, promoting inclusion, and combating issues such as cyberbullying, misinformation, and digital fraud. Education in digital citizenship aims to develop these skills, helping people, especially young people, navigate the digital world with awareness (Celot, Franceschetti, & Salamini, 2021; Cameron-Curry, 2014,)[2]. In summary, digital citizenship requires not only knowledge of technological tools but also a strong sense of ethics, responsibility, and respect for others. Digital citizenship means the combination of civic education and digital education, i.e. on the one hand, training in one's rights and duties as a citizen and on the other the awareness that the actions one takes online and offline have an impact in the present and future for oneself and for others. From a legal point of view, digital citizenship represents the set of rights and duties aimed at simplifying the relationship between businesses, citizens, and public administration using digital technologies. On the other hand, from an ethical point of view, digital citizenship means the ability to freely express oneself and one's ideas in the context of the infosphere

[2] See also the Council of Europe program: https://www.coe.int/en/web/education/dce-concept

by exercising freedom of choice and nurturing democratic life using digital technologies and within the digital environment.

The use of synthetic actors — digital or AI-generated entities that impersonate real or fictional persons — raises significant ethical concerns, especially in sensitive contexts such as education, public communication, and media. To ensure responsible and transparent use, it is necessary to establish rigorous standards that include:

1. **Informed Consent**
   It is essential that all individuals involved in the creation or representation through synthetic actors provide explicit and informed consent, understanding the purposes, modalities of use, and potential associated risks.

2. **Clear Labeling Requirement**
   All content produced using synthetic actors must be clearly labeled as such to avoid deception and ensure that users can recognize the artificial nature of the actor and the message.

3. **Independent Review for Sensitive Uses**
   In cases of use in particularly delicate areas — such as political information, education, healthcare, or justice — an independent third-party review is indispensable to assess the accuracy, safety, and social impact of the produced content.

Adopting these measures helps protect human dignity, prevent manipulation, and promote an ethical and informed digital culture.

To conclude this first section, it is important to highlight the ethical significance of digital citizenship. In this sense, it must be said that Ethics is a perspective on human fulfilment, the study and reflection on the principles, values, and norms that guide human beings toward the full realization of themselves and their potential. In this view, ethics is not merely a set of rules but a path that directs individuals toward what is considered "good" or "right," fostering harmonious development on both personal and social levels (Valera & Castilla, 2020; Riva, 2020).

From the perspective of ethics as human fulfilment, choices and actions are evaluated based on their ability to contribute to the well-being and growth of the individual, interpersonal relationships, and the community as a whole (Caltagirone, 2020). This vision sees ethical behaviour as a means to build an authentic existence, grounded in values such as justice, responsibility, respect, compassion, and honesty.

In summary, ethics as a perspective on human fulfilment considers ethical living not as an externally imposed goal, but as a journey toward the fullest development of one's human and relational capacities, in harmony with others and the world.

1.2 Digital Education and Digital Citizenship

Digital education is a discipline that is developing rapidly around the world, gradually becoming part of national education systems. It is a form of education that focuses on using digital technology to

help students learn to develop their skills in the field, in order to promote their growth and full personal development.

Thus, digital education is a process that involves the learning of digital skills, such as the use of computers, tablets and smartphones, but also the ethical training of the individual in his or her personal growth. Moreover, it can be used to teach children how to use digital devices. Not only that, it is also useful for instructing on the right way to surf the Internet, how to use social media and how to exploit digital resources. Digital education can also help people develop skills in critical thinking, relating to others, and self-understanding.

Digital education is key to building digital citizenship and democracy. It equips people to interact online responsibly and consciously, while also preparing them to participate in democratic life with critical thinking, respect, and openness. Digital education promotes abilities such as critical thinking, responsibility, privacy, and respect for digital rights, fighting misinformation, and active participation.

1.2.1 Educational strategies for developing digital citizenship and democracy skills

1. Digital Literacy and Critical Thinking: It is essential to teach students how to assess the reliability of information, distinguish trustworthy sources from fake news, and understand the mechanisms behind misinformation. Simulations and case studies can help develop these skills.

2. Privacy and Online Safety Education: Raise awareness of good practices to protect one's privacy and security (e.g., password management, privacy settings, recognizing phishing and online scams). Workshops and practical modules are helpful for learning these skills.

3. Civic Participation Experiences Online: Engage young people in digital participation projects, such as online discussion forums, civic surveys, or digital volunteer activities. These experiences teach the value of civic contribution and the importance of respectful, conscious participation.

4. Teaching Digital Ethical Values: Include activities in school programs that promote values such as respect, inclusion, and online responsibility. For example, lessons on cyberbullying and the importance of respectful language in digital interactions.

5. Collaboration and Online Conflict Resolution: Promote communication and conflict resolution skills to enable constructive management of differences online. Simulating scenarios of digital discussion or conflict provides practical skills for addressing disagreement.

6. Problem-Based Learning (PBL) Projects: Students work on real-world digital issues, such as privacy management, misinformation, or cybersecurity. This method encourages critical thinking and a sense of responsibility.

7. Education on Digital Rights and Duties: Help students understand their rights and duties online (right to privacy, security, freedom of expression, etc.). Providing resources to understand and respect digital laws and copyright is also key.

8. Digital Inclusion: Work to reduce the digital divide, ensuring that everyone has access to technologies and the ability to use them proficiently. Specific courses for less-digitized

populations can be essential for inclusive education (Agostini, 2024; Warchschauer, 2003; Molina & Mannino, 2016).

Implementing these strategies helps form aware digital citizens, prepared to contribute positively and responsibly to the digital society, strengthening democratic foundations through informed, respectful, and active participation (Boccacci, 2024; Rivoltella, 2020, 2024).

1.3 Opportunities for Using Artificially Generated Content in Education

Given the increasing reliance on digital platforms for education and building on digital education, the integration of AI-generated content (AIGC) **presents** new opportunities to enhance teaching and learning. The integration of GenAI and AIGC into education offers numerous advantages that traditional systems struggle to provide. These AI-driven technologies significantly enhance personalized learning, engagement, adaptability, and efficiency while fostering creativity and innovation.

1.3.1 Cultivating Curiosity and Personalized Learning

One of the most significant benefits of GenAI is its ability to foster curiosity in students, an essential factor for effective learning and academic growth. Traditional education often fails to promote a deep-seated curiosity, leading to disengagement and weaker student outcomes. If students lack the intrinsic motivation to explore new concepts, they may struggle to keep up with their peers. GenAI, however, can create engaging, interactive learning experiences that encourage curiosity and active participation, as supported by previous studies (Zeng et al., 2024).

Moreover, GenAI enhances personalized education by customizing content, learning pace, and complexity to meet each student's unique requirements, learning styles, and preferences. Digital tools like ChatGPT and similar GenAI technologies serve as effective digital tutors, offering instruction in various subjects to a diverse student population. Traditional education often fails to provide sufficient attention to each student, particularly those from different geographic or cultural backgrounds who may struggle to adapt to new teaching methods and learning environments (Rawal, 2020). In such scenarios, GenAI can bridge this gap by offering tailored strategies that help students acclimate to their new educational settings, ensuring a smoother and more effective learning experience.

1.3.2 Engagement and Interactivity

GenAI has the potential to revolutionize educational experiences by crafting immersive and interactive learning environments. This includes simulations, virtual reality scenarios, and gamified elements that enhance student engagement and eagerness to learn. Students may lose interest in specific subjects due to unengaging teaching methods or a lack of stimulating content. Traditional education systems often struggle to maintain student interest and facilitate effective learning (Huang et al., 2023). However, GenAI can rekindle student enthusiasm by introducing dynamic and captivating educational tools, leading to improved comprehension and retention of knowledge.

Additionally, GenAI enhances engagement by providing adaptive learning materials tailored to each student's interests and cognitive abilities. By leveraging AIGC, students can participate in virtual experiments, explore AI-powered educational scenarios, and interact with custom-generated content, fostering deeper understanding and practical application of knowledge (Chan & Hu, 2023). These

capabilities also support the development of higher-order cognitive skills, such as critical thinking, problem-solving, and creativity.

1.3.3 Adaptability and Individualized Support

A significant limitation of traditional education is that teachers can't always provide individualized attention to every student. Due to classroom constraints, many students' unique learning needs go unaddressed. GenAI, however, can dynamically adjust teaching methods based on a student's progress and achievements, offering personalized feedback, recommendations, and customized learning paths. This ensures that students receive the right level of support and challenge, promoting efficient learning and growth.

In a study conducted by Zhu et al. (2020) the effectiveness of an AI-powered feedback system was assessed in high school assignments related to climate studies. The results indicated that AI-driven feedback significantly improved students' scientific argumentation skills. By continuously monitoring student performance, GenAI can provide targeted interventions to strengthen weak areas while reinforcing mastered concepts, ultimately leading to more effective learning outcomes. Moreover, AI-generated assessments allow students to receive evaluations anytime and anywhere, ensuring continuous and accessible learning (Su et al., 2024).

1.3.4 Creativity and Innovation

Beyond structured learning, GenAI fosters creativity and innovation among students. It provides a platform for students to explore their artistic and inventive potential by generating unique artworks, music compositions, and written content. These AI-driven tools encourage students to think outside the box and engage in creative problem-solving. By offering personalized recommendations and creative prompts, GenAI helps students develop original ideas and express themselves in diverse ways.

Additionally, AIGC can facilitate collaborative learning experiences by offering tools and platforms that promote teamwork and knowledge-sharing. This is particularly crucial for developing communication skills and cooperative problem-solving abilities, which are essential in both academic and professional settings (Michalon & Camacho-Zuñiga, 2023). Artificial intelligence (AI) is transforming the educational landscape, making the role of teachers as critical mediators between students and digital technologies increasingly essential. Teachers are not merely users of AI tools, but guides who help students understand their potential, limitations, and ethical implications. For example, during a writing activity supported by a chatbot, the teacher can encourage students to reflect on the quality of the generated texts, compare them with authoritative sources, and rework them in their own words. In mathematics, a teacher might use an AI app to visualize complex problems but still prompt students to explain the solution process in their own terms. Furthermore, teachers can raise awareness about privacy protection by reminding students not to input personal data into automated systems. In this way, the teacher becomes a facilitator of conscious and responsible AI use, fostering critical, metacognitive, and digital skills that are essential for future citizenship.

1.4 Application of GenAI across Different Levels of Education

GenAI is transforming education by offering tailored learning experiences for students at different stages of their academic journey. From early childhood to university-level education, GenAI enhances engagement, fosters personalized learning, and provides innovative teaching methodologies.

1.4.1 Early Childhood Education

For young learners, interactive and engaging teaching methods are crucial in laying the foundation for cognitive and creative development. GenAI-powered intelligent boards can assist teachers in delivering educational materials while reducing reliance on physical resources. Audio-visual aids, known to positively impact children's creativity (Yazar & Arifoglu, 2012), can be seamlessly integrated into smart boards. AI-generated videos featuring catchy visuals and rhymes simplify fundamental subjects such as the alphabet, storytelling, and counting, making learning enjoyable.

Repetition is key in early education as it reinforces neural connections and aids retention (Montessori, 2017). GenAI can suggest innovative and engaging teaching strategies, such as interactive games and musical recitations. For instance, teachers can request AI-generated ideas for different kinds of activities (Mittal et al., 2024). Additionally, fostering collaboration between teachers and parents is essential in early education. A GenAI-powered application can bridge this gap by allowing teachers to upload progress reports, while AI suggests tailored home activities for parents. This ensures a cohesive learning experience that extends beyond the classroom, as suggested by Mittal et al. (2024). The growing use of digital tools and AI in children's daily lives presents new educational and health-related challenges. To support healthy development, screen time should be limited according to age: avoided entirely under age 2, kept to no more than one hour per day between ages 2 and 5, and individually regulated after age 6 to ensure it does not interfere with sleep, physical activity, or social interactions. It is essential to prioritize educational content, engage with children during screen use, and establish screen-free times, such as during meals or outdoor play. The introduction of AI also requires clear rules: children must understand that these tools are fallible and should be taught to use them critically, safely, and under supervision. Access should only be allowed when age-appropriate, with parental controls in place and special attention to data privacy. Parents and educators should model healthy behavior, encourage open dialogue, and guide young users toward responsible technology use. Only an integrated educational approach can ensure the benefits of digital tools without compromising children’s physical and mental well-being.

1.4.2 Middle and High School Education

As students progress into middle and high school, their learning needs shift towards more advanced theoretical and analytical concepts. GenAI can play a vital role in enhancing comprehension through personalized support and interactive learning tools. For example, Yixue Education’s intelligent tutoring system analyses students’ learning situations and performances by using AI and facial recognition to generate personalized tutoring content and strategies. Studies have shown that this system can improve the efficiency and effectiveness of students' learning, providing them with a better learning experience and support (Ouyang & Jiao, 2021).

GenAI assists students in improving their reading and writing skills by detecting grammatical and syntactic errors, thereby refining their language proficiency. AI-driven tutoring systems also provide personalized feedback, helping students enhance their writing abilities (Badoo-Anu & Ansah, 2023). Additionally, in subjects such as history and geography, GenAI facilitates understanding by generating immersive audio-visual content. The integration of metaverse technology further enhances this learning experience, allowing students to virtually explore historical sites, geographical landscapes, or scientific simulations (Lee & Hwang, 2022; Xu et al., 2024; Gaafar, 2021). This approach provides hands-on experiences that boost both engagement and comprehension (Cooper, 2023). GenAI also aids in complex STEM subjects by offering step-by-step solutions and visual representations of abstract mathematical and scientific concepts. It supports teachers by generating interactive quizzes, simulations, and real-time problem-solving guidance, making subjects like physics and chemistry more accessible.

1.4.3 Higher Education

At the university level, GenAI tools may serve as a powerful tool for understanding intricate and abstract academic topics. Criticisms and concerns notwithstanding, advanced AI models such as DALL-E and Midjourney generate detailed imagery that represents complex subjects that can be employed by teachers during classes, aiding in visual learning and making abstract concepts more comprehensible (Kohnke et al., 2023; Vartiainen & Tedre, 2023). This may prove particularly beneficial in fields such as medicine, engineering, and data science, where visual representation enhances conceptual understanding. GenAI may also foster peer-to-peer collaboration in both in-person and distance learning environments. AI-powered platforms adapt to individual inputs, offering personalized instruction and feedback that improve the effectiveness of virtual education (Doshi & Hauser, 2023). By promoting active participation in group discussions, AI-driven learning platforms may enhance collective learning experiences. Furthermore, GenAI may play a crucial role in making education more inclusive for individuals with disabilities. It offers real-time text-to-speech and speech-to-text capabilities, ensuring accessibility for students with visual or auditory impairments. This technology provides personalized support, enabling students of all abilities to engage fully in their educational experiences (Zhang et al., 2024).

1.4.4. Impact on Teachers' Roles

GenAI may enhance efficiency in education by automating various tasks such as generating worksheets, tutorials, evaluations, and feedback. This has the opportunity of reducing the administrative workload for teachers, allowing them to dedicate more time to personalized instruction and student engagement. AI-driven assessments provide fine-grained insights into students' thought processes, enabling educators to evaluate higher-order skills such as critical thinking, reflective reasoning, and problem-solving abilities (Exintaris et al., 2023).

AIGC platforms equip educators with advanced data analysis capabilities, allowing them to identify learning gaps and refine instructional strategies to enhance both student engagement and academic performance (Celik et al., 2023; Liu & Wang, 2021). Furthermore, one of AIGC's key strengths is automation, which streamlines various teaching-related tasks such as grading, lesson planning, and quiz generation. This significantly reduces educators' workload, enabling them to focus more on personalized instruction and student interaction (Asthana & Hazela, 2021; Celik et al., 2023; Liu & Wang, 2021). Additionally, the incorporation of natural language processing and machine learning in

AIGC systems facilitates real-time feedback, progress monitoring, and more interactive and meaningful engagement with students by introducing elements of gamification, simulations and more interactive activities (Asthana & Hazela, 2021; Celik et al., 2023; Liu & Wang, 2021).

As a consequence, the adoption of AIGC is also reshaping the role of educators, giving greater priority to their role of learning facilitators and mentors who guide students through higher-order thinking and creative problem-solving (Lameras & Arnab, 2023; Liu & Wang, 2021; Shen & Su, 2021). This transformation necessitates the development of new skills in AIGC integration, instructional design, and mentorship, highlighting the growing need for continuous professional development (Lameras & Arnab, 2023; Liu & Wang, 2021).

Beyond redefining traditional roles, AIGC empowers educators with innovative tools such as data analytics, adaptive learning technologies, and AI-driven tutoring systems, enabling a more personalized and informed approach to teaching (Asthana & Hazela, 2021; Bisen et al., 2022; Celik et al., 2023). Furthermore, the integration of AIGC-powered virtual and augmented reality enhances the learning experience by creating immersive environments that push the boundaries of conventional education (Bisen et al., 2022; Dilmurod & Fazliddin, 2022).

1.4.5 Effects of Using AIGC on Learning Experience and Learning Outcomes

Artificially Intelligent Generated Content (AIGC) plays a significant role in enhancing the learning environment by actively monitoring student progress and identifying areas that require additional support (Asthana & Hazela, 2021; Celik et al., 2023). The ability of AIGC to personalize learning enhances student comprehension, encourages self-paced learning, and helps learners achieve their full potential (Bisen et al., 2022; Celik et al., 2023; Dilmurod & Fazliddin, 2022; Lameras & Arnab, 2023). By allowing students to select topics, control their learning pace, and engage in preferred activities, AIGC promotes autonomy and ownership over the learning process, accommodating diverse learning styles.

Beyond conventional methods, AIGC incorporates elements of gamification, simulations, and interactivity to increase student engagement (Popenici & Kerr, 2024; Huang et al., 2022). Gamification strategies, such as integrating game-like elements into educational activities, enhance motivation and participation. Simulations provide immersive experiences that enable students to apply theoretical knowledge in practical contexts, fostering deeper understanding. Interactivity through AIGC platforms facilitates collaboration, discussion, and active participation, cultivating a dynamic and engaging learning environment. Virtual learning spaces further exemplify AIGC's impact on engagement, enabling students to explore concepts interactively and collaborate with peers beyond traditional classroom constraints (Chaiyarak et al., 2022; Khurana et al., 2020).

The effectiveness of AIGC-generated instructional content compared to human-made materials remains a topic of ongoing research. Leiker, Gyllen, et al. (2023) found no significant difference in learning experience when students engaged with an AI avatar instead of a human teacher. However, their study compared a single human instructor in a talking head video to a distinct AI-generated avatar, limiting generalizability (Beege et al., 2022). The debate continues over whether AI-generated videos can match human-made teaching materials in relatability and emotional impact. While human instructors bring personal experiences and emotional nuance that enhance content engagement

(Schneider et al., 2022; Tack et al., 2022), research suggests that AI-generated narratives can reduce counterarguing and foster story-consistent emotional responses (Chu & Liu, 2023). However, labelling content as AI-generated may trigger scepticism, reducing its effectiveness due to decreased narrative immersion.

Studies on AI-generated content's readability and logical coherence present mixed findings. Markowitz et al. (2024) argue that AI-generated text tends to be more analytic and descriptive but also less readable than human-generated content, potentially impacting students' comprehension of educational videos. Although AI-generated materials may offer structured and logically coherent content, students may find human-created narratives more familiar and engaging due to natural imperfections that aid retention (Beege et al., 2022). Furthermore, while AI-generated videos may minimize human distractions such as speech inconsistencies and physical movements, these very imperfections might enhance attentiveness, thus supporting better internalization of learning material (Sondermann et al., 2024).

In terms of learning outcomes, research has yielded mixed results. Some studies suggest that AI-generated teaching materials can be as effective, if not more so, than traditional instruction. Schroeder et al. (2013) concluded that digital pedagogical agents have small but positive effects on learning. More recent research supports this, indicating that AI-generated content can enhance knowledge retention and transfer (Leddo et al., 2021; Pi et al., 2022). However, Leiker et al. (2023) found no statistically significant difference in learning gains between students who watched AI-generated videos and those who viewed human-led videos. Their findings contribute to the growing body of literature exploring AI-generated synthetic videos compared to traditionally produced instructional content. While AI-generated synthetic videos mimic traditional talking-head instructional formats, pedagogical agents and avatars have traditionally been designed to improve learning outcomes by incorporating gestures and emotional cues (e.g., Horovitz & Mayer, 2021).

Pi et al. (2022) further reinforce the potential of virtual instructors, showing that AI-generated instructors in video lectures can facilitate learning on par with human instructors. Their findings align with the equivalence principle, which suggests that virtual and human instructors can be equally effective in video-based learning. Notably, they observed even better transfer outcomes when learners engaged with AI-generated videos featuring a synthesized instructor image and voice, emphasizing the potential of AI-generated teaching materials in digital education.

1.4.6 Ethical and Security Challenges of GenAI in Education

Training GenAI models for educational content, such as quizzes and reading materials, raises concerns regarding copyright infringement and plagiarism. These models risk generating responses that violate intellectual property laws and may store user inputs as training data without consent. Applications across various educational domains, including early education, medical education, and skill development, must address these risks through compliance with copyright laws, transparent content usage policies, and user awareness initiatives (Hayes, 2023; Ren et al., 2024).

GenAI models, particularly Generative Adversarial Networks (GANs), pose challenges in control and predictability, making them unreliable for structured learning environments. Unchecked AI outputs can mislead students, particularly young learners, and generate misinformation. Ensuring AI

explainability, rigorous testing, and adherence to curriculum guidelines can mitigate such risks (Sun et al., 2022). Moreover, GenAI requires extensive datasets, yet they lack true creativity, often synthesizing existing information rather than producing genuinely novel content. This limitation affects personalization in education and long-term content value (Jiang et al., 2021).

Bias in GenAI models is another critical issue, as seen in cases like self-driving cars misidentifying darker-skinned individuals (Wilson et al., 2019). Addressing bias requires diverse training datasets, continuous monitoring, and fairness-driven AI development. Tools like IBM Watson OpenScale and AI Fairness 360 can help detect and mitigate biases. Transparent AI mechanisms and educational programs for students and teachers are essential for fostering responsible AI use (Yu et al., 2024).

Lastly, it is important to acknowledge the risks that research has found to be connected to the use of digital technologies in education in general. The use of digital technologies in education is associated with several risks, including reduced reading and writing proficiency due to reliance on typing rather than handwriting (Tan et al., 2013), impaired memory from shallow information processing (Craik & Lockhart, 1972; Craik & Tulving, 1975; Sparrow, Liu, & Wegner, 2011), increased distraction and multitasking that lower academic performance (Bowman, Levine, Waite, & Dendron, 2010; Ellis, Daniels, & Jauregui, 2010; Hembrooke & Gay, 2003), and even digital addiction among students (Baek & Park, 2013). As a consequence, the choice of using technology in education should be pondered carefully while considering both benefits and risks.

In summary, while GenAI holds immense potential in education, it requires strict regulatory frameworks, bias elimination strategies, and ethical safeguards to ensure its responsible deployment.

2. The Impact and Legal Implications of GenAI on Different Population Segments

This section aims to map population groups involved in the use of GenAI to illustrate the socio-political impact of these technologies. Segmenting the population to measure deepfakes' impact on the legal sphere involves identifying specific groups that may be affected differently. However, this alone is insufficient for a full understanding of the legal consequences. Despite growing concerns about the trustworthiness of GenAI, there is a lack of available data and literature on valid segmentation criteria to measure their trustworthiness in the legal sphere. A comprehensive European dataset on the main legal consequences of GenAI does not yet exist. For the legal field, segmentation groups should, at a minimum, include lawyers, judges, clerks, and public administrators.

A deeper understanding of the intersection of GenAI, the legal sphere, social media, and population segmentation is crucial for developing effective strategies to mitigate risks and maximize the benefits of this technology for legal purposes and future regulatory development. The lack of dialogue across disciplines is highlighted as a critical issue that also limits the legal field.

A first approach to understanding legal accountability and the applicable consequences of GenAI begins with an analysis of their diverse applications across social media platforms. These applications include:

- **Content creation:** This involves deepfakes (video and audio), synthetic images, and AI-generated avatars used in marketing, influencer bots, or fake profiles.
- **Content manipulation:** This is achieved through image and video editing, and voice cloning.
- **User behaviour simulation:** This involves bot accounts and GAN-based profile photos.
- **Advertising and influence campaigns:** Targeted propaganda and fake testimonials, with electoral campaigns being a notable example.

While empirical work on deepfakes has increased since 2020, legal research, particularly when it applies empirical methods, remains scarce. A systematic literature review by Berri and Just (2024) identified a key issue: the limited empirical evidence to support current regulatory needs. For example, there is a lack of research on how deepfakes are created, used, and spread, or their potential societal impacts, making it unclear if voiced concerns align with actual problems. It is noteworthy that only Article 50.4 of the EU AI Act directly addresses this phenomenon, posing a challenge for regulatory development by Member States. The SOLARIS project is working to understand these empirical research gaps and to identify new cross-disciplinary research models for deepfakes.

Social-science research has predominantly focused on the societal impact of deepfake disinformation and the psychological harm from deepfake pornography. Behavioural indicators and other quantitative and qualitative measures could provide more data on how GenAI affects social media and users, which could inform public policies and awareness campaigns. Studies have increasingly focused on the use of deepfakes in political campaigns, an issue seen as eroding democratic systems. Literature review and consulted research conducting cross-citation analysis (Berri & Just, 2024) reveals that studies rarely

cite each other across disciplinary boundaries, indicating a predominantly disciplinary, rather than interdisciplinary, nature. This is clearly a limitation for drafting effective legal instruments.

2.1 Beyond social media: Segmentation studies for trustworthiness and legal accountability

Given the gap in studies on GenAI effects on the legal sphere, it is crucial to highlight emerging efforts to design trustworthy public digital services at the EU level and elsewhere. National authorities have responded with various regulatory solutions, including information correction, content removal, and criminal sanctions, which can be justified when carefully tailored to protect legitimate interests (R. Helm & H. Nasu, 2021).

A systematic breakdown of legal issues, which the SOLARIS project is undertaking, can be used to design empirical models and segmentation criteria to address the effects of GenAI in the legal sphere. These issues include:

- **Obstruction of justice:** GenAI can create fake evidence, potentially undermining trust in legal proceedings.
- **Identity theft and reputation damage:** Deepfakes can be used to create false videos, harming reputations and facilitating identity theft, affecting anyone, not just celebrities.
- **Election interference:** This is a well-developed area of research showing deepfakes can cause confusion, undermine institutional trust, and exacerbate political polarization.
- **Defamation:** GenAI can generate fake content leading to defamation and libel cases, often related to hate speech.
- **Intellectual Property:** GenAI raises questions about authorship, ownership, and copyright infringement.
- **Consumer Protection:** The use of generative and predictive AI in marketing for segmentation and persona development contrasts with the need to protect consumer rights.
- **Regulatory challenges:** The rapid evolution of deepfake technology and the lack of a uniform global regulatory framework make it difficult for policymakers to keep pace.

A more comprehensive understanding of deepfakes' legal impact requires further research in collaboration with social sciences.

2.2 Vulnerable socio-economic groups

This section presents a literature review on socio-economic groups vulnerable to deepfake targeting and manipulation within the European context. Instead of analysing established socio-economic groups, this section categorizes them based on the specific harms posed by deepfakes: **victims**, **targets**, and the **unaware**. This approach recognizes that the challenges of

deepfakes necessitate a rethinking of how technology-disadvantaged groups are defined, moving beyond traditional barriers to include populations intentionally targeted by technology.

Over the past two decades, AI has become a global phenomenon, but its pervasive nature raises questions about equitable access and benefits. As Kluge and de Oliveira (2021) note, there are ethical and socio-economic concerns that AI-powered machines may make vulnerable populations superfluous. Technological innovations bringing about disadvantages for specific groups are not a new trend; the digital divide, for example, shows how digital tools can reproduce existing discrimination (Lythreatis et al., 2021). The challenge posed by deepfakes, however, requires a new taxonomy for identifying technology-disadvantaged groups. It becomes critical to determine who is most likely to have their features appropriated to generate fake content, who the intended audience is, and how this audience overlaps with those who not only view but also trust the deepfake. While some research points to an "AI-penalty," or an inherent distrust of algorithmic outputs (Lee & Moshirnia, 2024), this defense diminishes as AI becomes more precise and its inputs less detectable, supported by the network economics framework. It is therefore an urgent matter to rethink which individuals are most exposed to this technology and to identify the causal paths of this relationship.

2.2.1 Victims

While deepfake incidents are widely reported in the media, there has been limited empirical research to define at-risk socio-economic groups. Based on existing literature and reports, three major groups are most at risk of misrepresentation: women and girls, ethnic minorities, and local political figures.

- **Women:** Since the emergence of deepfake images in 2017, academics and commentators have argued that deepfake pornography is a deeply misogynistic practice that disproportionately harms women, akin to non-consensual pornography (Harwell, 2018; Öhman, 2020; van der Nagel, 2020; Rini & Cohen, 2022; Badham, 2024). In 2019, cybersecurity firm Deeptrace estimated that 96% of over 14,000 deepfake videos in circulation were pornographic, with targets almost exclusively women (Adjer et al., 2019). The proliferation of free "nudify" tools fuelled a 550% increase in deepfake pornographic images between 2019 and 2023 (Home Security Heroes 2023). While initially targeting public figures, the technology's accessibility means anyone may be vulnerable (Sensity, 2024; Sippy et al., 2024). A survey across 10 countries found that 2.2% of participants were personally victimized by non-consensual deepfake pornography (Umbach et al., 2024), with increasing reports of young women being targeted in schools and universities (Beazley & Touma, 2024; Mackenzie & Choi, 2024). Victims often experience severe psychological trauma (Laffier & Rehman, 2023). Deepfakes are also used as tools of control to suppress activism, and can cause harm by changing the discursive context around a person, forcing them to react to fabricated images (Rini & Cohen, 2022). The use of deepfake pornography for CSAM, exploitation, and sextortion is also a significant issue (Sippy et al., 2024).
- **Ethnic minorities and marginalized groups:** These groups are particularly vulnerable due to existing societal biases and the influence of racial tensions on political discourse (Concha, 2023; Skene, 2025). The effectiveness of deepfake detection tools is impacted by social biases, with error rates 10.7% higher for individuals with darker skin tones (Trinh & Liu, 2021). Additionally, evidence shows that people have difficulty identifying deepfakes of

individuals from other racial groups (Lovato et al., 2024). More broadly, the erosion of public trust caused by deepfakes may lead to the dismissal of videos documenting real racial injustices as fake, undermining accountability (Pfefferkorn, 2021).

- **Local public figures:** While deepfakes targeting high-profile figures are often quickly debunked, a growing concern is "microfakes" that target lesser-known individuals like local officials and community leaders. These deepfakes often go undetected (Ascott, 2020). Although the immediate political impact of a single microfake may seem limited, their widespread use poses a serious, granular threat to democracy. Experimental studies confirm that fake videos alleging political scandals can significantly reduce favourable attitudes and voting intentions, even when the deepfakes are of low quality (Dan, 2025). A particularly concerning development is the use of deepfake pornography as a political tool to undermine female public figures (Maddocks, 2020).

2.2.2 Targets

The dissemination of deepfakes is not uniform; specific groups are more susceptible to manipulative content due to cognitive, social, and structural factors. These groups become strategic targets for actors aiming to influence political opinions, manipulate social narratives, or impact institutional trust.

- **Elderly people:** Older adults are particularly vulnerable to misinformation, in part due to limited access to diverse information sources and lower digital literacy (Moore & Hancock, 2022). They are more susceptible to information overload (Vivion et al., 2024), and a RAND Corporation survey found that those over 55 are especially vulnerable to online misinformation when it comes in unfamiliar formats like deepfakes (Huguet et al., 2024). They tend to rely on superficial cues to judge credibility, which is ineffective against highly realistic manipulations (Tang et al., 2024). This vulnerability is further compounded by broader cognitive and behavioural differences in how older adults process online information (Xing et al., 2020).

- **Young people:** Young individuals, particularly in their early teenage years, often lack the knowledge to critically evaluate online content. They may accept information uncritically, especially if it aligns with their beliefs or comes from familiar sources. Research indicates that adolescents frequently rely on superficial cues like the presence of images or post popularity, rather than analysing source credibility (Wineburg & McGrew, 2016). This challenges the assumption that "digital natives" possess inherent critical digital literacy (Robb, 2020). The social nature of their online interactions further exacerbates this vulnerability (Hassoun et al., 2023).

- **Minorities and socioeconomically disadvantaged groups:** These groups often face systemic barriers that limit their access to accurate information, making them vulnerable to disinformation. They may be more inclined to trust information that appears to offer solutions to their struggles without critical evaluation. During the COVID-19 pandemic, studies showed that migrant and ethnic minorities relied heavily on social media, increasing their exposure to misinformation (Goldsmith et al., 2022). Research also shows that misinformation can circulate rapidly within their densely connected communities (Karimi et al., 2024), and that a

perception of minority status can increase susceptibility to conspiracy beliefs (Gundersen et al., 2023).

- **Citizens in politically sensitive countries:** These individuals are often more susceptible to disinformation campaigns due to a history of media manipulation, limited access to diverse information, and the strategic interests of both domestic and foreign actors. Tactics include the use of fake news websites and deepfake videos to sow conflict, undermine trust, and manipulate electoral outcomes, as seen during the European Parliament elections (Reuters, 2024). The constant exposure to conflicting information can lead to political polarization and a decline in democratic engagement (Bennett & Livingston, 2018).
- **People with high conspiracist ideation:** Individuals with a strong predisposition to conspiracy beliefs are particularly attractive targets for disinformation. Their cognitive patterns, such as confirmation bias (Lewandowsky, Oberauer, & Gignac, 2012), lead them to accept information that reinforces their pre-existing narratives. They may even interpret corrections as further proof of a conspiracy (Franks, Bangerter, & Bauer, 2013). This psychological mechanism makes them highly receptive to emotionally charged deepfakes and fake news that fit their worldview.

2.3 Psychological Approaches to Segmentation

This section draws from a broader body of literature on the psychological determinants of susceptibility to misinformation and complements it with findings specifically on deepfakes. It addresses the roles of demographic, motivational, and cognitive variables.

- **Demographic variables:** Previous literature suggests that demographic variables, particularly age and social media use, are important in the context of misinformation (van der Linden, 2022). Existing studies show that older individuals are generally more susceptible to misinformation due to cognitive decline and lower digital literacy (Chen et al., 2023; van der Linden, 2022). This finding also applies to deepfakes, as vulnerability to deepfakes increases with age (Doss et al., 2023), and age is significantly associated with judgments regarding deepfake trustworthiness (Plohl et al., 2024).
- **Motivational variables:** These include political orientation, conspiracy beliefs, and trust. Extreme political orientations are consistently linked to higher susceptibility to misinformation.
- **Cognitive variables:** Education, media literacy, reflective thinking, and "bullshit receptivity" are all associated with an individual's ability to discern and resist misinformation. Higher education and media literacy generally act as protective factors, while higher "bullshit receptivity" is a strong predictor of finding deepfake content more trustworthy.

2.3.1 Segmentation supported by psychological approaches

This section draws from a broad body of literature on the psychological determinants of susceptibility to misinformation, complementing it with findings from studies that have specifically investigated individuals' vulnerability to deepfakes. It is structured to first examine the role of sociodemographic variables, followed by an analysis of motivational and cognitive factors.

2.3.2 Demographic Variables

Previous literature has established that demographic variables, particularly age and social media use, are significant in the context of misinformation (van der Linden, 2022). While susceptibility can vary, existing studies suggest that older individuals are generally more vulnerable, which is partly attributed to cognitive decline and lower digital literacy (Chen et al., 2023; van der Linden, 2022). This finding also extends to deepfakes. For example, Doss and colleagues (2023) concluded that vulnerability to science deepfakes increases with age across all stakeholder groups. Similarly, our study found that age was significantly and positively associated with individuals' judgments regarding the trustworthiness of deepfakes, in both their content and presentation (Plohl et al., 2024).

Social media use is another variable linked to misinformation susceptibility. Individuals who use social media more frequently may fall victim to the **illusory truth phenomenon**, where repeated claims are more likely to be judged as true (Chen et al., 2023; van der Linden, 2022). Similar findings have emerged in the context of deepfakes, with frequent social media users tending to be less sceptical (Ahmed, 2023). Our deliverable [FR1] provides a more detailed look into this association. Specifically, we found that using social media as a news source was positively associated with the perceived trustworthiness of content, but not with the perceived trustworthiness of presentation (Plohl et al., 2024). This suggests that repeated exposure may make individuals more vulnerable to questionable arguments but may not affect their ability to discern manipulated from authentic videos.

It is worth noting that findings on the relationship between gender and misinformation susceptibility are inconsistent. While some studies suggest men are more likely to trust misinformation, others report the opposite (Chen et al., 2023). In our study, there were no significant differences in the perceived trustworthiness of deepfake content or presentation between men and women (Plohl et al., 2024).

2.3.3 Motivational Variables

Drawing on theories like **motivated reasoning**, which posits that decisions are often based on predetermined goals rather than a factual evaluation of evidence (Kunda, 1990), researchers have identified individual variables that can motivate a person to believe misinformation. These include political orientation, belief in conspiracy theories, and trust.

A more extreme and right-wing political orientation has consistently been linked to higher susceptibility to misinformation, even on non-political topics (Chen et al., 2023; van der Linden, 2022). More recent literature emphasizes that political orientation intersects with other factors, with a study by Lawson and Kakkar (2022) showing that misinformation sharing is largely driven by conservatives with low conscientiousness. This pattern also appears in deepfake literature. A study by Sutterlin and colleagues (2023) reported a strong negative association between a participant's conservatism and their ability to recognize fabricated videos. Our study in UC1 (Plohl et al., 2024) added a layer of complexity, showing that conservatism was positively associated with the perceived trustworthiness of deepfake content but not with the trustworthiness of its presentation, demonstrating an informational bias but no difference in technical recognition skills.

Chen and colleagues (2023) explain that a belief in conspiracy theories is positively correlated with believing and sharing misinformation, as misinformation often reinforces these beliefs (van der

Linden, 2022). However, research on this in the context of deepfakes is scarce and shows mixed results. For instance, Nas and de Kleijn (2024) found that individuals with higher conspiracy thinking were better at distinguishing deepfakes from authentic videos. Our results [FR2], obtained during a validation study, showed no association between conspiracy beliefs and the perceived trustworthiness of deepfakes. Therefore, the role of conspiracy mentality in the perception of deepfakes remains unclear and is likely highly context-specific; it may increase general distrust but could also increase perceived trustworthiness when deepfakes support a particular conspiracy theory.

Trust is a complex variable in the context of misinformation. Generally defined as the intention to accept vulnerability based on positive expectations of another person or institution (Dirks & Ferrin, 2002), it can refer to interpersonal or institutional trust. Some forms of trust are protective, such as trust in scientists in the context of COVID-19 misinformation (Roozenbeek et al., 2020). However, blind trust in unreliable media can increase susceptibility (Chen et al., 2023). In our study, trust in media was significantly and positively associated with the perceived trustworthiness of deepfake content but not its presentation (Plohl et al., 2024). Interestingly, our unpublished data from Use Case 1 [FR3] showed that general trust (as a disposition) was positively associated with the perceived trustworthiness of presentation but not content. Ultimately, trust acts as a double-edged sword: it is essential for credible communication but can also make individuals vulnerable to deception when unwarranted (O'Brien et al., 2021).

2.3.4 Cognitive Variables

In addition to demographic and motivational variables, cognitive abilities and related variables play a crucial role. The **inattention account** posits that the constant bombardment of information, coupled with limited time and resources, hinders individuals' ability to critically reflect on content. Previous research consistently shows that education, media literacy, reflective thinking, and "bullshit receptivity" are all associated with misinformation processing (Roozenbeek et al., 2020; van der Linden, 2022).

While higher education is generally linked to better discernment of true and false news (van der Linden, 2022), studies on deepfakes often focus on specific education and knowledge. For instance, Hwang et al. (2021) reported that media literacy education may reduce the effects of deepfake disinformation. In UC1 [FR4], we found no significant association between general education and the perceived trustworthiness of deepfake content or presentation. However, we did find a significant negative association between media literacy and the trustworthiness of content. Conversely, self-reported knowledge about deepfakes reduced the perceived trustworthiness of the presentation but not the content itself (Plohl et al., 2024).

Reflective thinking, which involves a deliberate and analytical approach to information processing, has also been linked to better misinformation detection (van der Linden, 2022). Our study (Plohl et al., 2024) found that individuals with higher reflectiveness perceived the content of deepfakes as significantly less trustworthy, although reflective thinking was not associated with their perception of the presentation.

Finally, "bullshit receptivity," a construct describing the tendency to ascribe profundity to randomly generated sentences, has been positively associated with perceptions of fake news accuracy

(Pennycook & Rand, 2019). In our study, this variable was the strongest correlate of the perceived trustworthiness of deepfake content, with more receptive individuals finding the content more trustworthy. However, bullshit receptivity was not associated with the perceived trustworthiness of a deepfake's presentation (Plohl et al., 2024).

2.4 A Theoretical and Statistical Review of Segmentation Approaches in Fake News and Deepfake Impact Analysis

This review critically synthesizes existing literature on segmentation strategies for fake news and deepfakes. It examines psychological segmentation approaches—rooted in individual differences in cognition, motivation, emotion, and personality—alongside statistical methodologies like logistic regression, latent class analysis, and structural equation modeling. This integrated lens provides a comprehensive theoretical and empirical foundation for understanding how misinformation operates across population strata.

The inquiry also highlights significant gaps in the current literature, including a scarcity of research focused explicitly on GAN-generated misinformation, a limited integration of multimodal segmentation dimensions, and underexplored implications for democratic processes. To address these gaps, the review advocates for a more holistic, multidimensional approach that transcends disciplinary boundaries and leverages advances in data science, psychology, and political communication.

2.4.1 Psychological Approaches to Segmentation

A nuanced understanding of misinformation susceptibility requires exploring the psychological mechanisms that govern how individuals engage with information. Segmentation based on these psychological dimensions is essential for identifying vulnerable subgroups.

- **Cognitive Traits and Information Processing:** Cognitive style, particularly analytical thinking, is a central determinant of misinformation discernment. Pennycook and Rand (2019) showed that individuals with higher scores on the Cognitive Reflection Test (CRT) were less susceptible to false information. Martel et al. (2020) further supported this, finding that cognitive reflection facilitates accurate news classification. These results suggest that cognitive segmentation can inform targeted interventions.
- **Personality Traits and Fake News Engagement:** Personality-based segmentation, using the Big Five framework, is also fruitful. Wolverton and Stevens (2020) found that individuals high in extraversion and low in conscientiousness were more likely to disseminate fake news. Sindermann et al. (2020) identified openness and conscientiousness as positively associated with fake news detection.
- **Motivational Factors and Belief in Misinformation:** Motivation to process information systematically is another critical factor. Calvillo et al. (2021) found that individuals with a high need for cognition were less inclined to accept misinformation. The Elaboration Likelihood Model (Petty & Cacioppo, 1986) provides a theoretical basis for this, distinguishing between central and peripheral routes of processing.

- **Emotional Responses and Content Engagement:** Emotion-driven engagement plays a pivotal role in misinformation diffusion. Vosoughi et al. (2018) showed that false content spreads faster due to its emotional appeal, and Brady et al. (2017) demonstrated that moral-emotional language amplifies diffusion. Emotional susceptibility is thus a salient dimension for segmentation.
- **Trust and Source Credibility Perception:** Trust in an information source is another vector influencing belief. Nedelcu (2021) found that individuals heavily weigh source credibility, while Buchanan and Benson (2019) reported that trust in social media platforms enhances the perceived credibility of content shared on them.
- **Demographic Variables: Age and Media Trust:** Age-related differences are robustly associated with misinformation engagement. Brashier and Schacter (2020) argued that older adults' higher rate of sharing fake news is due to lower digital literacy and more homogeneous social networks.

2.4.2 Statistical Approaches to Segmentation

The psychological dimensions discussed above gain empirical robustness when translated into statistically formalized segmentation frameworks. These methods quantify latent traits and behaviours, creating predictive models that validate theoretical assumptions.

- **Demographic and Socioeconomic Segmentation:** Logistic regression provides foundational insights into age, education, and political orientation as predictors of misinformation behaviour. Guess et al. (2019) found that users over 65 were nearly seven times more likely to share fake news. Allcott and Gentzkow (2017) demonstrated the influence of political alignment.
- **Behavioural Segmentation:** This approach identifies user clusters based on interaction patterns. Nelson and Taneja (2018) found that a small group of highly active users accounted for most of the traffic to fake news sites. Grinberg et al. (2019) showed that political engagement, not orientation, was the strongest predictor of fake news sharing.
- **Psychographic Segmentation:** This directly maps psychological constructs to statistical clusters. Pennycook and Rand (2019) used CRT scores to partition participants by cognitive style, finding that higher scores aligned with reduced susceptibility. Martel et al. (2020) used factor analysis and SEM to uncover distinct cognitive profiles.
- **Geographic Segmentation:** This highlights spatial heterogeneity. Badawy et al. (2018) used machine learning to segment Twitter users by state, revealing regional variation in susceptibility. Bovet and Makse (2019) showed that misinformation flows are modulated by geographic clustering.
- **Technological Segmentation:** Segmenting users by technological preferences clarifies the role of media architecture. Fletcher et al. (2018) found that consumers of different platforms had different levels of exposure to misinformation. Cinelli et al. (2020) showed that echo chambers vary significantly by platform.

- **Cross-Platform Behaviour Segmentation:** Integrative models, such as those by Gottfried (2024), use cluster analysis to segment users by their behaviour across multiple platforms, revealing distinct user types with differing susceptibility levels.

Despite the advances from statistical models, key gaps remain. First, GAN-generated deepfakes are underrepresented in segmentation studies. Second, there is a scarcity of comprehensive models that simultaneously incorporate psychographic, behavioural, and demographic variables. Third, most research is geographically constrained to Western contexts, limiting generalizability. Finally, the societal ramifications of misinformation, such as the erosion of civic trust, are not sufficiently linked to empirical segmentation frameworks.

2.4.3 Statistical Modelling Approaches for Studying the Impacts of GenAI

The exponential growth of misinformation, accelerated by GenAI, poses a threat to public discourse. As these technologies produce increasingly persuasive fabricated content, there is a pressing need for rigorous analytical frameworks to understand the spread of fake news and individual susceptibility. This chapter examines three key statistical techniques: logistic regression, latent class analysis (LCA), and structural equation modelling (SEM).

- **Logistic Regression:** This serves as a foundational tool for modelling binary outcomes, such as the likelihood of an individual engaging with misinformation. It helps identify significant risk factors and estimate their influence.
- **Latent Class Analysis (LCA):** This provides a probabilistic framework for uncovering unobserved heterogeneity within a population, identifying and characterizing latent subgroups with different vulnerabilities. It offers nuanced insights into how distinct audience segments engage with news content.
- **Structural Equation Modelling (SEM):** SEM allows researchers to simultaneously examine complex interrelationships among observed and latent constructs. It is well-suited for testing theoretical models that link cognitive traits, political ideology, media literacy, and affective responses to misinformation susceptibility.

Together, these three methodologies provide a comprehensive toolkit for empirical research on fake news, offering both granular and integrative perspectives on its spread, reception, and social consequences. They enable scholars and practitioners to develop more precise, data-driven strategies to mitigate the effects of algorithmically amplified misinformation.

2.4.4 The Proposed Statistical Models

In the empirical study of misinformation and fake news, particularly those generated or amplified by GenAI, the deployment of advanced statistical modelling techniques is critical for unpacking the intricate interplay of individual, cognitive, and sociotechnical factors. These approaches enable researchers to move beyond descriptive patterns to inferential insights, revealing latent structures, causal pathways, and probabilistic predictors of susceptibility. The most frequently used techniques are logistic regression, latent class analysis (LCA), and structural equation modelling (SEM). Each method offers distinct analytic affordances for capturing different dimensions of the misinformation phenomenon.

3. Description of ANT (Actor–Network Theory) Methodology for Understanding GenAI for Good

This section presents a literature review of the ANT methodology, also used in D2.2, its application to GenAI and how such an approach might be used to promote the use of these technologies for social good, particularly regarding deepfakes. The section is structured into three sub-sections. The first sub-section provides a broad theoretical introduction to ANT methodology, highlighting its unique approach to non-human actors. The second sub-section focuses on literature explaining the application of ANT methodology to GenAI, identifying the key social actors at play in the production and dissemination of AI-generated content. The third sub-section then focuses on the notion of using AI for social good and how this concept relates to ANT. Below are suggested readings for each section.

3.1 ANT methodology

Closely associated with the work of sociologists Bruno Latour (2005, 2017), Michel Callon (1986), and John Law (1992), actor-network theory (ANT) offers a radical perspective on how our social reality is constructed and maintained. Marking a significant departure from traditional sociological theory, ANT scales down the influence of abstract concepts such as rigid social structures and social forces and instead conceptualises any social activity as a constantly shifting network of concrete relationships between different social actors. Within ANT, the notion of a "social actor" does not solely refer to a human being but further encompasses a broad spectrum of entities, including objects, animals, texts, technologies, and institutions. Where traditional sociology privileges human intentions and actions, ANT considers human and non-human social actors as equally important to the dynamics and activities of the network. As such, machines, institutions, and objects can impact our social environment as much as human beings. This means that ANT does not privilege human beings and, instead, conceptualises all of these different social actors as equal actors continually interact with one another in a flat non-hierarchical network. As the interactions between these social actors change, so too does the composition and dynamics of the overall network change. As such, the boundaries of this network are fluid and never fixed, and so any social actor, be they human or non-human, may take on a more active role and come to instigate significant change within the larger network. In emphasizing the fluidity and material reality of social relationships and further recognising the role of non-human social actors, ANT allows for a more nuanced description of the explicitly social function of GenAI technologies and their increasingly influential role of AI-generated communications (e.g., deepfakes) within our socio-political environments.

To elaborate, the ANT approach emphasizes the materiality of networks, thus proposing that socio-political ideas, knowledge, and values are established and maintained by the material interactions between those social actors involved in a network. As a notable example, theorists such as Latour, Callon, and Law have argued that scientific facts are not simply fundamental truths revealed through rational inquiry but rather are social products that have been materially constructed through the interactions between different social actors involved in scientific research. This includes human actors (e.g., scientists, engineers, administrators) and non-human actors such as objects (e.g., scientific equipment), processes (e.g., established methodologies), and institutions (e.g., research centres). Furthermore, these socially constructed scientific facts do not simply exist as abstract concepts in people's minds but rather always exist in a variety of material forms (e.g., publications, patents,

databases, conference presentations, learned skills). Additionally, the dissemination of these facts is also mediated by a larger material network of objects, people, and institutions (e.g., postal systems). As an example, Latour has argued that Louis Pasteur did not "discover" germs but, rather, he aligned different social actors (e.g., laboratory instruments, microbes, scientific papers, and institutes) into a network that gave legitimacy to germ theory (Latour, 1988). This is not to argue against the legitimacy of scientific methods specifically, but rather, to highlight the role of social interactions in the construction and perpetuation of facts and knowledge.

ANT has its limitations, however, which Latour, Callon, and Law have often highlighted. Notably, the methodology has received criticism for overemphasising the agency of non-human social actors (Law & Hassard, 1999) and for neglecting the influence of established hierarchies and power inequalities by insisting that all social actors are equivalent (Winner, 1993). Furthermore, because ANT analysis insists upon a vast and fluid network of social interactions, it is often a laborious effort to map all relevant social actors and interactions involved in a specific case study. As such, some have argued that ANT is an impractical methodology, particularly for analysing large-scale social structures (Law & Hassard, 1999).

That said, ANT's unique perspective allows us to more appropriately account for the increasingly significant role that GenAI technologies play within our social interactions and, furthermore, allows us to expand our discussions to consider the roles of various other social actors. To consider the role of GenAI in the production and dissemination of socio-political ideas, then, it is necessary to step back from the immediate interaction itself (e.g., a person viewing a deepfake) to instead consider the far wider material network of social actors that frames this interaction (e.g., media organizations, policymakers, government institutions). The material network of GenAI technologies may be vast and ever-changing, such that we can only provide an approximation.

3.2 3.2 ANT for AI-Generated Content

Analysing AI technologies such as deepfakes through Actor-Network Theory (ANT) offers a powerful lens to understand their complex socio-technical entanglements. ANT allows us to go beyond the deterministic view of technology by treating AI systems not just as tools, but as active participants (actors) within extended networks. These networks include human agents - developers, users, regulators - along with non-human entities such as algorithms, datasets, and online platforms. The advantage of ANT, as Morton (2024) exemplifies in his analysis of ChatGPT, lies in its ability to track how these different actors mutually shape actions and contribute to emergent outcomes, revealing the distributed nature of agency and accountability in AI-driven contexts. By tracing the interactions between these actors, ANT provides a richer understanding of the socio-technical forces driving the development, diffusion, and impact of AI, challenging simplistic narratives of technological progress or human control. Moreover, this approach can reveal the subtle ways in which seemingly neutral technologies can embody and perpetuate existing social biases, as evidenced by the skewed datasets often used to train AI models. Rather than focusing solely on human agency or technological determinism, ANT encourages us to explore the interaction between different actors - developers, algorithms, datasets, platforms, and users - treating them all as equally significant contributors to the network.

Moreover, an ANT approach applies *to any network constructed at any node of a social discourse.* As an example, the extension of the ANT approach to legal discourses on AI offers crucial insights into the regulation of these technologies. Traditional legal frameworks often struggle to address the complex interactions within AI networks, particularly the changing power dynamics and liability issues. By mapping the relationships between AI systems, legal actors, and stakeholders, ANT can illuminate gaps and tensions within existing regulatory structures. For example, when considering the proliferation of deepfakes, ANT can reveal how platform algorithms, user behaviour, and evolving legal definitions of authenticity interact to influence the spread and impact of synthetic media, as highlighted by Dasilva et al. (2021) in their study on Twitter deepfakes. This approach highlights the need for adaptable legal frameworks that recognise the fluidity and complexity of AI networks, moving beyond simplistic notions of cause and effect to embrace a more systemic understanding of responsibility.

The *spread of* deepfakes introduces a specific level of complexity into the network. Social media platforms, for instance, act as critical 'crossing points' where content is filtered and amplified. The algorithms that govern these platforms play an active role in determining which deepfakes gain visibility and which are suppressed. This dynamic raises questions about power and agency: are platform designers responsible for the dissemination of harmful content, or are the algorithms themselves exercising a form of agency? ANT allows us to avoid these binary distinctions, focusing instead on how human intentions and technological possibilities combine to produce specific outcomes. For example, as Hajli et al. (2022) explained in their study on social bots and misinformation, automated accounts—created by someone or an institution with a given purpose—often amplify deepfake content, creating feedback loops that further reinforce its presence in public discourse.

Detection technologies add another dimension to this network. Tools designed to identify deepfakes, such as fingerprint estimation networks, interact with generative models in a kind of arms race. As detection methods improve, so do the techniques to circumvent them. This iterative relationship underlines the fluidity of the networks studied through ANT: actors continually adapt in response to one another. Moreover, the survey instruments themselves are not neutral: they reflect the priorities and assumptions of their creators. By tracing these connections, ANT can reveal how technical decisions made during development reverberate externally to influence social perceptions of authenticity and trust.

We should also add that *the integration of ANT with semiotics***,** from which Bruno Latour famously drew his early models, can further enhance our understanding of AI as a cultural and communicative phenomenon. Semiotics examines how AI systems generate, manipulate, and interpret meaning, revealing how these technologies reshape our understanding of truth, authenticity, and trust. As Hepp et al. (2023) argue in their analysis of communicative AI systems, these technologies are transforming media ecosystems and challenging traditional notions of human communication. By combining ANT with semiotics, we can analyse how AI actors participate in semiosis, the process of meaning-making, both intentionally and unintentionally. For example, a deepfake video not only presents a manipulated image, but also engages with existing cultural codes and visual conventions, creating a complex semiotic message that can be difficult to decipher. More specifically, semiotic studies on AI emphasise its role as a 'technology of the fake', capable of producing synthetic content that blurs the

boundaries between reality and simulation. Ian Goodfellow's seminal work on GANs is an example of this: the architecture is based on an antagonistic relationship between a generator (which creates synthetic data) and a discriminator (which assesses its authenticity). This dynamic mirrors the semiotic process of signification, where meaning emerges through the interaction between signifiers and their interpretation. Poulsen's semiotic analysis of deepfake software further explores how these tools appropriate cultural signifiers, such as facial expressions and gestures, to create synthetic images that carry rich layers of meaning. By embedding these signifiers in multimodal contexts (e.g. video with speech or motion), deepfakes exploit the semiotic potential of faces as powerful bearers of identity and emotion.

3.3 Using ANT to promote social good

While the discourse around AI often focuses on dangers stemming from potential misuse, such as threats to democracy, liberty, and privacy, threats to peace and safety, and threats to work and livelihoods, one cannot deny that AI also has the potential to revolutionise healthcare (e.g,, early diagnosis and screening, drug discovery) and other sectors (e.g., translation, climate modelling, interactive learning) in a way that benefits our society (Federspiel et al., 2023). Such duality also applies to AI-generated deepfakes, in which a person in an existing image or video is replaced with someone else (Somoray & Miller, 2023). On the one hand, deepfakes pose a significant risk to our society due to putting pressure on journalists who need to distinguish real and fake news, endangering national security by spreading propaganda, and creating cybersecurity threats. On the other hand, they may also have positive uses in various areas, including the film industry, educational media, digital communications, and healthcare (Westerlund, 2019). In line with this, it is becoming increasingly clear that AI and specific technologies relying on it should not be treated in a black-and-white manner. Instead, we, as a society, should attempt to maximize their benefits and minimize their risks. To better understand how to approach this complex goal, this section reviews the existing literature on using AI, including deepfakes, for social good, and proposes actor-network theory as an additional approach that may lead to new insights related to the interplay between technology, human actors, and societal structures.

The potential positive impact of AI is often explored under the umbrella term of AI for social good (also known as AI4SG), which is rapidly gaining momentum (Akula et al., 2021). While specific perspectives on what constitutes social good (and hence AI for social good) differ, AI4SG generally encompasses building tools and solutions that help address the world’s most pressing challenges and deliver positive social impact in accordance with the United Nations’ sustainable development goals (Tomašev et al., 2020). The latter includes 17 goals, including achieving good health and well-being, providing quality education, and facilitating climate action (United Nations, n. d.). As these goals are rather diverse, so are the resulting AI4SG endeavours, which include building models that forecast clinical manifestations, helping individuals avoid phishing, and promoting student retention, to name a few (Akula et al., 2021). While the literature is still very scarce, similar principles can also be applied to deepfakes, which may, for instance, be used to increase motivation and attention among learners, enrich interactive role-playing, and create new learning materials (Danry et al., 2022).

While different authors propose different specific characteristics of AI4SG, there is generally a consensus regarding the key features of such technology. Several researchers agree that such technology should, as a starting point, be designed to respect fundamental human rights. As explained

by Aizenberg and van den Hoven (2020), these include *dignity* (i.e., ensuring that individuals do not experience a state of helplessness, insignificance, or humiliation), *freedom* (i.e., respecting individual autonomy, freedom of expression and information, and their right to privacy), *equality* (i.e., ensuring non-discrimination based on various grounds), and solidarity (i.e., combating social exclusion). The same authors use these fundamental rights to derive the central values that need to be respected by AI systems, such as non-discrimination, privacy, transparency, and safety (Aizenberg & van den Hoven, 2020). Relatively similar AI4SG values can be found in other works as well. Umbrello and van de Poel (2021), for example, highlight the importance of respecting human autonomy, preventing harm, fairness, and explicability. Similarly, Akula et al (2021), who introduced the term ethical AI for social good, argue that the main components are explainability and interpretability, privacy protection, data safeguards, intervening only in a way that respects users' independence, fairness, and lack of discrimination, adaptability and user-friendliness, and verification and validation before deployment. We argue that many of these values are also crucial in the context of deepfakes, in particular preventing harm and ensuring transparency, with both being addressed in the AI Act, which states that harmful AI-based manipulation and deception pose an unacceptable risk, and that deployers of deepfakes should disclose that the content has been artificially generated or manipulated (European Commission, 2025).

Besides broader regulations, other ways to ensure that AI contributes to social good include developers' compliance with guidelines on the overall use of AI technology, applications, and data handling, participatory design, and value-sensitive design. First, the guidelines, which are described in detail in the article by Tomašev and colleagues (2020), encourage developers to design inclusive and accessible AI applications, establish and maintain trust to overcome barriers, and ensure secure data processing. In contrast, participatory design is a bottom-up approach that includes consulting the relevant stakeholders in the design of AI solutions via a range of empirical methods, such as surveys, interviews, focus groups, participant observation, and participatory prototyping (Aizenberg & van den Hoven, 2020). The involvement of diverse stakeholders may lead to systems that are more beneficial to individuals and contribute to higher trust (Zhang et al., 2023).

Participatory design is often considered a component of value-based design, a broader approach to designing new technologies that takes values into account (Umbrello & van de Poel, 2021). As opposed to traditional approaches to design, which largely focus on technological capabilities, value-based design marks a shift to proactive considerations of the societal context in which the technology is embedded and how societal needs and values can be translated into technology design (Aizenberg & van den Hoven, 2020). Value-based design relies on various methods, summarized by Friedman and colleagues (2013) as "tripartite methodology". This approach integrates three types of investigations, namely conceptual investigations (i.e., identifying the stakeholders and values implicated by the technological artifact in the given context), empirical investigations (i.e., exploring stakeholders' needs, views, and experience in relation to the technology and the values it implicates), and technical investigations (i.e., implementing and evaluating technical solutions that support the values and norms elicited from the conceptual and empirical investigations). Interestingly, while there are several published articles about employing value-based design to develop AI solutions, to our knowledge, only one study has used such methodology in the context of deepfakes. Specifically, Maia et al (2024) investigated whether deepfakes may be used to counter misinformation ethically and effectively. They involved 11 stakeholders from various countries and domains to ensure ethical

implementation of AI in fact-checking. They concluded that "*By involving stakeholders, implementing transparency measures, and establishing robust ethical guidelines, it is possible to harness the positive aspects of this innovation while mitigating its potential drawbacks*" (p. 13).

We argue that value-based design may be complemented and further improved with ANT to promote the use of AI, including deepfakes, for social good. First, ANT provides a systematic approach to identifying the relevant stakeholders involved in developing, deploying, and regulating AI and deepfakes, which is a focal ingredient of the tripartite methodology described above. While value-based design emphasizes stakeholder involvement, ANT surpasses this by also mapping the complex relationships between human and non-human actors, shedding light on how their interactions shape the technology's social impact. This can help ensure that all crucial actors, including those that may otherwise be overlooked, are included in the design and governance process. Second, ANT highlights how ethical principles, such as transparency and harm prevention, must be negotiated within a socio-technical network. Establishing ethical AI guidelines is not enough; their successful implementation depends on interconnected actors, including regulators who enforce standards, media institutions that implement these measures, and users who develop media literacy. By tracing and reinforcing these networks, ANT provides a framework for ensuring that AI for social good is not just a theoretical goal but a practical reality. Lastly, ANT emphasizes the evolving nature of socio-technical networks, making it particularly useful for understanding the long-term implications of AI for social good. Technologies like deepfakes are not static; their uses, risks, and societal responses shift over time as new actors enter the network and relationships between existing actors change. ANT allows for continuous monitoring and adaptation by recognizing that ethical AI governance must be an ongoing, iterative process. This perspective can help policymakers, developers, and other stakeholders remain responsive to emerging challenges while reinforcing the positive applications of AI-generated content.

4. Enhancing Ethical Digital Citizenship: The Role of AI-Generated Social Actors

The identification of synthetically generated actors is a crucial step in understanding the social impact of AI. These actors, often indistinguishable from real ones, pose new challenges in recognition, trust, and interaction processes. Their analysis allows for the segmentation of the population into groups based on age, digital literacy, and media exposure, revealing different levels of vulnerability or awareness. This approach enables a more nuanced study of AI, highlighting how it influences perceptions and behaviours. Segmentation, therefore, is not just a statistical tool but an interpretative key for assessing AI's societal impact. In this context, synthetic identification becomes an integral part of computational social sciences.

This section presents a literature review of the ethical considerations that need to be considered when identifying what kinds of synthetically generated social actors should be used for pro-democratic AI-generated content. The section is structured into three sub-sections, each focusing on different kinds of synthetically generated social actors. This includes (i) deepfakes of deceased people, (ii) deepfakes of living people, and (iii) non-existent AI avatars. Each section summarises the major ethical arguments and ideas in the literature around these different kinds of synthetically generated social actors.

4.1 Deepfakes of the deceased

The use of GenAI to digitally resurrect deceased celebrities and public figures has become increasingly popular, particularly in film, television and advertising. However, this practice is highly controversial, raising significant ethical concerns surrounding consent, exploitation, and- personal autonomy, as well as the emotional and psychological impact (Bozdağ, 2024; Bassano & Cerutti, 2024; Nash, 2022).

One of the primary ethical issues around deepfakes of the deceased relates to consent and posthumous rights. While often framed as tributes to the deceased, these digital resurrections might constitute a form of commercial exploitation, particularly in the entertainment industry. Though a deceased individual's legal representatives (e.g., surviving family, estate) may grant permission for their likeness to be used, the person themselves cannot provide explicit consent. As such, these digital resurrections might publicly link the individual's image to a media production (e.g., film, TV show) or a commercial product that they did not or would not endorse in their lifetime. Not only does this practice present a possible infringement of an individual's posthumous rights over the use of their own image but it also risks artificially distorting or changing their previously established public image or persona. Media productions and commercial products are themselves linked to and/or promote ideological messages. As such, using a digitally resurrected individual to represent and/or endorse these productions and products further draws artificial associations between the deceased individual and these socio-political messages, thus altering their perceived identity in a way that the deceased did not or would not consent to.

This issue of ideological association raises other ethical issues when using deepfakes of the deceased within an explicitly socio-political context, and particularly with historical figures, as SOLARIS is exploring in relation to UC3. Using digital resurrections of historical figures for promoting specific

contemporary political ideas (e.g., ethical digital citizenship, democratic values, equality, and diversity) that they did not or would not endorse in life constitutes an exploitation of their image for political gain and distorts their own personality and political opinions. For example, using a deepfake of the notoriously anti-immigration British politician Enoch Powell to promote immigration and multi-ethnic societies would seem to be a blatant distortion of the man's known political views and an exploitation of his image for political purposes that go against his own. Furthermore, such digital resurrections may be seen as so obvious as to draw attention to historical contradictions and thus undermine the credibility of the messages they seek to promote. Such efforts might very well backfire and be seen as an attempt to deceive or manipulate the public sewing furthering distrust in the media environment.

This issue is further exacerbated if we are to transplant these historical figures from their socio-political context in the past to our own contemporary context. Even if a historical figure was known to promote a specific socio-political issue in life (e.g., immigration), the discourse and debate around this issue may have become significantly different in our contemporary context. Not only would this misrepresent the figure's position by recontextualising their documented views on a specific socio-political issue, but it also misrepresents the issue itself as something universal or as an unchanging problem that has been the same throughout history. Such deepfakes of historical figures would seem to distort public understanding of history and oversimplify political issues as eternal debates isolated from context, rather than things that are eternally changing with changes in context or whatever.

While using deepfakes of recently deceased figures might avoid this disconnect between historical contexts, there is a risk of causing emotional and psychological harm to the individual's surviving relatives, their colleagues, and the communities of which they were a part. Seeing a highly realistic digital recreation of a deceased loved one could be distressing and might even disrupt the grieving process. While there is no conclusive evidence, some experts (Hollanek & Nowaczyk-Basińska, 2024; Rovetta & Valentini, 2025) suggest that such hyper-realistic representations could prolong grief or cause psychological harm, especially if the individual is depicted in ways that feel inappropriate or exploitative.

In using deepfakes to promote democratic values, SOLARIS should be cautious about the individual figures recreated and aware of the ethical implications regarding these figures' known political positions and original socio-political contexts, as well as the potential emotional impact of their digital resurrection.

4.2 Deepfakes of the living person

The creation of deepfakes featuring living people raises similar ethical concerns as those related to deepfakes of the deceased, notably issues of consent, association with unwarranted ideological messages, and misrepresentation of an individual's political views. If approached carefully, however, we could obtain consent from participants and ensure that the deepfakes produced align with their own political views and values to eschew these issues. Indeed, deepfakes of living people have been used in positive ways (Kietzmann, 2021). For example, deepfakes can enable highly personalised political messaging that is aimed at the viewer's specific context and also allows for easy translation of these messages into multiple languages. Such technology might enable people to publicly communicate in unprecedented ways. For example, following the 2024 Pakistan election in which

supporters of imprisoned former president Imran Khan received a significant share of the vote, a deepfake of Khan was able to deliver a victory speech while the man himself remained in prison and was unable to issue public communications. However, while we may be able to obtain consent from living individuals and ensure that deepfakes of them align with their own political views, deepfakes do present further unique and distinct harms for living people (Rini & Cohen, 2022; Öhman, 2022; Crippen, 2023).

Notably, there are unique and distinct emotional and psychological harms for those depicted in deepfakes. While primarily discussing the impact of deepfake pornography, Rini and Cohen (2022) argue that deepfakes may cause illocutionary harm, a form of harm that occurs when individuals are forced to respond to fabricated content about themselves. This can include public figures being compelled to deny deepfake videos that depict them saying or doing things they never did. Even if a deepfake is widely disbelieved or openly fake, simply having to deny it publicly or address its existence is a form of harm, as it forces individuals to engage in speech acts against their will.

Even if individuals agree to be used for deepfake political messaging, they may face consequences for doing so in their personal and professional lives. These fabricated representations could expose these individuals to public criticism, reputational damage, and even personal harm in extreme cases. They may be open to criticisms of manipulation and deception. These risks of deception and backlash are especially heightened when deepfakes involve contemporary socio-political figures, as their associations carry immediate and often polarizing implications.

For SOLARIS to publicly show deepfakes of living individuals for political messaging, explicit and detailed consent must be obtained, and certain steps must be taken to ensure that these individuals will not face physical or psychological harm as a result (UC1).

4.3 Non-existent AI Avatars

AI integration in education has been widely explored in recent research, particularly regarding AI-generated avatars and their role in digital learning environments. The existing literature highlights both the potential benefits and ethical concerns associated with these technologies. Novelli and Sandri (2024) emphasize the need for transparency in AI-generated social actors within online democratic spaces. Their work aligns with the discussion on AI avatars in education by advocating for ethical frameworks to regulate synthetic actors, ensuring authenticity in digital interactions. Similarly, Emejulu and McGregor (2019) argue for a critical digital citizenship approach that acknowledges the socio-political implications of AI-driven technologies. These perspectives highlight the importance of ensuring AI avatars operate within ethical boundaries, fostering trust and mitigating misinformation. Several studies have explored the role of AI avatars in enhancing student engagement and personalized learning. Rodriguez and K (2023) present a comprehensive analysis of AI avatars in higher education, discussing their ability to personalize learning experiences, provide real-time feedback, and improve inclusivity. Their work builds on prior research by Pataranutaporn et al. (2021), which investigates how AI-generated characters support personalized learning and student well-being. Similarly, Vallis et al. (2023) examine student perceptions of AI avatars in teaching business ethics, finding that while AI avatars offer interactive experiences, scepticism remains regarding their effectiveness compared to human instructors. Beyond education, AI-generated avatars have been studied in business and consumer interactions. Nalivaikė and Miliukaitė (2024) explore

how AI-generated avatars influence consumer trust, brand recognition, and loyalty. Their findings suggest that AI avatars can promote subconscious emotional and cognitive trust, at times outperforming human actors with consistent and specifically engineered emotional expressions. This parallels their use in education to establish engagement and credibility in digital learning environments. Despite the advantages of AI avatars, ethical concerns remain a central discussion point in the literature. Rodriguez and K (2023) outline issues such as privacy risks, algorithmic bias, and data security, echoing concerns raised by researchers like Leiker et al. (2023), who investigate the implications of GenAI in learning videos. Moreover, Harvard University's AI ethics research has emphasized the necessity of clear guidelines to ensure data privacy and bias mitigation in AI-driven education systems.

5. Co-creation methodology and citizen engagement for the Sustainable Development Goals

This section introduces the co-creation methodology and citizen engagement approach employed to address the Sustainable Development Goals (SDGs). It outlines how citizens were directly involved in creating and analysing synthetically generated content,

5.1 Visual semiotics approaches in UC3

Use Case 3 (UC3) of the SOLARIS project aimed to qualitatively investigate the civic and communicative potential of GenAI through the participatory creation and analysis of video content based on synthetic actors. in UC3 three online co-creation workshops were organised, each lasting two hours, with a total of 44 participants selected from more than 700 initial registrations. Unlike Use Case 1, which focused on quantitatively measuring the cognitive and attitudinal effects of political deepfakes, UC3 was grounded in citizen science principles, and also adopted a semiotic and processual perspective: the objective was not only to observe/evaluate effects of persuasion or misinformation, but to understand how artificial texts[3] are constructed, what dimensions make them credible, engaging, or off-putting, and how participants attribute meaning to them.

Within the adopted methodological framework, semiotics was understood as the science of meaning-making forms and their conditions of production and interpretation. We did not consider deepfakes merely as technological objects or in terms of deception, but as texts endowed with expressive form and discursive structure, capable of activating inferences, expectations, emotional reactions, and ethical evaluations in their viewers. In this sense, our attention focused on the textual complexity of synthetic videos. Each video, in semiotic terms, can be analysed as a text articulated on various levels:

- **Discursive level:** Assumes that any "story," even a still image, is about something, projecting characters into a specific space and time.
- **Narrative level:** Concerns what the characters are doing, and what changes or evolves in the storyline.
- **Enunciative level:** Includes the relationship between the sender and receiver, contracts of truthfulness, and the framing regime (e.g., fiction, testimony, or a hybrid).
- **Axiological level:** Expresses the implicit or explicit values conveyed by the text (e.g., truth, authority, empathy, transparency).

Furthermore, in the realm of visual texts, visual semiotics becomes relevant. As outlined by Polidoro (2008), visual semiotics is a branch of semiotics that addresses the production and interpretation of visual and audiovisual texts, with a particular focus on perception. It traditionally divides into two areas: figurative semiotics, analysing meaning derived from object and scene recognition, and plastic semiotics, concentrating on the significance conveyed by visual configurations such as shapes, colours, textures, lighting, and rhythm. This approach enables the identification of multiple layers of meaning involved in evaluating a visual text, as well as the criteria that support its plausibility or

[3] In semiotics, the term “text” refers to a cultural element—or a group of cultural elements—of any kind, in any format, on any medium, and in any language, which is isolated and reconstructed for the purpose of analysis. In other words, the notion of text is methodological, not empirical.

authenticity. The interpretation of an image does not rely solely on perceptual or realistic elements but is mediated by cultural and cognitive mechanisms. Visual literacy develops over time through familiarity with communicative genres, aesthetic codes, and narrative conventions. In this context, semiotic analysis explores the signals that render an image believable, plausible, or alternatively, disturbing and unacceptable—especially when dealing with artificially generated content. A key analytical dimension concerns the distinction between enunciation and utterance, as theorized by Greimas and Courtés (1986). The utterance is the completed message with internal coherence and reference to content. Enunciation, by contrast, encompasses the implicit or explicit traces that indicate the message was produced by a subject in a given context. The credibility of a message depends not only on what is said, but on how, by whom, and under what conditions. Often, perceived truthfulness arises from the coherence between enunciation and utterance; however, dissonance between the two levels may generate uncertainty or rejection.

Another central aspect concerns the relationship between verbal, visual, and auditory components of synthetic audiovisual texts. AI-generated content—such as deepfakes—is structured by the simultaneous interplay of multiple modes. Semiotics has shown that verbal components play an anchoring role: they guide the reading of an image or video, orient interpretation, and influence the regime of truthfulness. Captions, titles, logos, subtitles, introductory statements, or ethical disclaimers are all elements that shape meaning, and their presence or absence can radically alter a text's reception. A third analytical level involves the isotopic consistency of visual texts. Semiotics distinguishes between plastic isotopies—coherence in perceptual parameters such as lighting, texture, and visual mimesis—and figurative isotopies—coherence between depicted elements and cultural universes. Content becomes more credible when it achieves semantic redundancy between what it shows and what the audience expects to find in that context. When this coherence fails—through background anomalies, detail incongruities, or mismatches between body language and verbal content—suspicion or artificiality is perceived. In the case of synthetic content aimed at civic, ethical, or educational goals, the semiotic approach offers essential tools for designing texts that are not only formally realistic but also culturally and enunciatively credible. Only through a conscious balancing of these various layers of meaning can content be developed that is both innovative and responsible, capable of establishing a relationship of trust with the public.

5.2 Semiotic analysis of experimental data in UC3

UC3 set out to explore the possibility of designing positive deepfakes—synthetic videos aimed not at manipulation or deceptive simulation but at public information, historical memory, health promotion, scientific education, and active citizenship. A key feature of the UC3 methodology was the direct involvement of citizens in co-creating the content. Far from being mere observers, participants were asked to actively reflect on the goals, limitations, and potential of the proposed videos, formulating hypotheses, suggestions, critiques, and alternative proposals. This phase had two major effects. First, it generated a reflexive dialogue between content producers and users, making the implicit sense contracts within the texts explicit. Second, it allowed for real-time feedback: participants' decoded forms, detected incoherencies, interpreted communicative intentions, and distinguished between technical realism and narrative plausibility. Co-creation thus functioned as a device of semiotic negotiation, bringing to light the codes, values, and expectations that structure digital reception.

The eight videos developed for UC3 were designed to systematically yet creatively and originally test a set of semiotic and discursive variables relevant to how audiences perceive synthetic content. The scripts were conceived around three major civic themes, linked to three SDGs: historical memory and SDG 5 Gender equality; climate crisis and SDG 13 Climate Action; and mental health SDG 3: Good Health and well-being. Each theme was paired with a different representational mode and experimental variable configuration.

A. **Theme: Historical Memory - Marie Curie** (SDG 5: Gender equality). Three videos portrayed Marie Curie not as a celebratory figure but as an authoritative witness reflecting on the role of women in science.

    - **Video 1:** A deepfake reconstruction of Marie Curie, seated indoors in an early 20th-century café, narrates her career struggles in the first person. The face is synthetic, the voice neutral.
    - **Video 2:** Marie Curie appears livelier, speaks in the third person, and is set against a black background.
    - **Video 3:** Marie Curie's granddaughter speaks through the voice and face of her grandmother. The character is dressed in a more modern way, the scene is a modern scientific laboratory, and the character speaks at length about women and STEM.

B. **Theme: Climate Crisis - Amina** (SDG 13: Climate Action). Three videos address climate migration through the testimony of a young woman, Amina, who recounts leaving her homeland near Lake Chad due to desertification.

    - **Video 4:** Deepfake images show the consequences of climate change with a noticeably non-neutral voice-over explaining them.
    - **Video 5:** Amina tells her story with a non-authentic face, as a disclaimer informs. Climate change is now related to a personal experience. The voice from Video 4 is revealed to belong to Amina, and the background is a digitally created library.
    - **Video 6:** Amina's face is blurred—despite already being synthetic. This deliberate estrangement effect tests the tolerance for ambiguity between identity protection and credibility loss.

C. **Theme: Mental Health - Casey** (SDG 3: Good Health and well-being). Two videos explore psychological therapy mediated by an AI avatar, through the testimony of a young man, Casey.

    - **Video 7:** Casey narrates in the first person how an AI avatar modeled on his image helped him overcome social anxiety. The camera angle is face-on.
    - **Video 8:** The spoken content is identical, but the camera angle changes to a back-facing shot. The character is shown from behind to test the threshold of credibility without a face.

During the co-design phase with workshop participants, these eight scripts were discussed. Each was intended to test at least two of the eight semiotic variables defined during preparation.

(a) Famous vs. anonymous person;

(b) Realistic vs. decontextualized setting;

(c) Monologue vs. dialogue;

(d) Detail vs. totality;

(e) Blurred vs. reconstructed synthetic face;

(f) First vs. third person;

(g) Artificial landscape vs. artificial person;

(h) Serious vs. spectacular context.

For sustainability reasons, we focused on variables (a), (b), (e), (f), and (g), prioritizing the three themes. Workshop analysis highlighted several key insights. The body is always fully semiotic: even a nape, hair texture, or breathing posture contributes to realism. Their absence, as with Casey seen from behind, signals artificiality and undermines content. Voice is also a crucial vector of meaning: prosody, rhythm, ambient noise, and emotional intonation are key to credibility. TikTok-style text-to-speech voices were perceived as cold and inappropriate. Younger participants seemed to assess videos primarily by technical quality (lip sync, frame rate), treating it as a non-negotiable threshold for plausibility. It might be suggested that older participants were more attuned to discourse, its purposes, and its ethical-political implications.

Unsurprisingly, participants' comments suggested that context functions as a framework for meaning-making: institutions such as museums, schools, social media, and health clinics embody distinct semiotic regimes. A deepfake might be pedagogically useful in an educational setting, yet perceived as misleading or disturbing in a promotional context.

Several strong semiotic lines emerged from the workshops. **Enunciative transparency** (info banners, persistent disclaimers) increases acceptability, especially when deepfake use is declared and contextualized. **Narrative coherence** between avatar appearance, visual context, and speech content is key to avoiding uncanny effects (e.g., a smiling face recounting trauma is disturbing). In the discussion, participants expressed positive opinions about the possibility of using collective avatars (a group deepfake rather than a single actor) and interactive formats (Q&A modules or chatbots). Synthetic avatars may be better suited to rendering the invisible (remote past, inaccessible environments, abstract concepts) than to replacing living figures in caregiving or authoritative roles.

One major experimental limitation of UC3 lies in the isolation of deepfakes from the dynamic ecosystem of real-world circulation. A deepfake's meaning and impact depend on surrounding discourses, platforms, comments, sharing frames, and viewing devices. Each deepfake thus inscribes itself in a network of texts, codes, and communicative practices that shape its reception, purpose, value, and credibility. In UC3, videos were shown in standardized, decontextualized conditions, limiting analysis of the pragmatic and rhetorical mechanisms that usually determine a synthetic text's effectiveness. A fuller evaluation will require situating them within real communicative ecosystems to understand their intertextuality, cause-and-effect dynamics, and plasticity in public use.

The collected observations form a basis for a **taxonomy** distinguishing: (1) different genres of positive deepfakes (testimonial, expository, evocative, educational, interactive); and (2) diverse reception modes and perceived effects based on semiotic variables, audience type, and usage context. This dual taxonomy—textual and interpretive—can provide an operational tool for designing deepfakes that are truly **AI for Good**: socially, educationally, and ethically driven while respecting user expectations and competencies. This interpretation remains provisional; its value lies in mapping the most relevant semiotic nodes and outlining future design pathways.

6. Conclusion

This deliverable has demonstrated that generative artificial intelligence (GenAI) constitutes both an opportunity and a profound challenge for democratic engagement. On the one hand, its applications in education, participatory practices, and digital inclusion show significant potential for fostering creativity, critical thinking, and civic participation. On the other hand, risks such as bias, disinformation, manipulation, and violations of dignity and privacy underscore the urgent need for safeguards. The evidence presented makes clear that vulnerable groups—women, minorities, youth, the elderly, and socio-economically disadvantaged communities—are disproportionately affected, highlighting the importance of equity as a guiding principle for policy and practice.

From a theoretical perspective, the adoption of Actor–Network Theory has illuminated how GenAI functions not merely as a technical instrument but as a social actor that redistributes meaning, values, and power. The co-creation methodology piloted within the project has confirmed that democratic legitimacy cannot be ensured through regulation alone but requires citizen participation in shaping the very technologies that influence public life. Similarly, the semiotic taxonomy of synthetic media developed here offers both conceptual clarity and practical tools for fostering transparency, accountability, and trust in AI-mediated communication.

The overarching conclusion of this work is that the benefits of GenAI for education, inclusion, and democratic life can only be realized through a multidimensional strategy that integrates legal, technical, ethical, and participatory measures. Clear labelling and transparency rules, investments in digital literacy, interdisciplinary research, and co-design processes with citizens represent indispensable components of such a strategy. Only by embedding GenAI within a framework of responsibility, justice, and human dignity can societies harness its transformative potential while protecting democratic values and the common good..

References

Aizenberg, E. & Van Den Hoven, J. (2020). Designing for human rights in AI. *Big Data and Society*, *7*(2), 2053951720949566. https://doi.org/10.1177/2053951720949566

Aizenberg, E. & Van Den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*, *7*(2), 20539517

Abowd, J. M, (2017). How will statistical agencies operate when all data are private? *Journal of Privacy and Confidentiality*, *7*(3), 1–15. doi:10.29012/jpc.v7i3.404

Akula, R., & Garibay, I. (2021). Ethical AI for social good. In C. Stephanidis, M. Kurosu, J. Chen, G. Fragomeni, N. Streitz, S. Konomi, H. Degen, & S. Ntoa (Eds.), *HCI International 2021 Conference Proceedings – Late breaking papers: Multimodality, extended reality, and artificial intelligence* (pp. 369–380). Springer International Publishing.

Albanian Institute of Science (AIS). Open Data Albania – Political Financing Reports.

Allen, C., Smit, I., & Wallach, W. (2005). "Artificial morality: Top-down, bottom-up, and hybrid approaches", *Ethics and Information Technology*, *7*(3), 149–155. doi:10.1007/s10676-006-0004-4

Allen, C., Varner, G., & Zinser, J. (2000). "Prolegomena to any future artificial moral agent", *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261. doi:10.1080/09528130050111428

Amoroso, D., & Tamburrini, G. (2018). "The ethical and legal case against autonomy in weapons systems", *Global Jurist*, 18(1): art. 20170012. doi:10.1515/gj-2017-0012

Anderson, J., Rainie, L., &, Luchsinger, A. (2018). *Artificial Intelligence and the Future of Humans*, Washington, DC: Pew Research Center.

Anderson, M., & Leigh Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent, *AI Magazine*, *28*(4), 15–26.

Aneesh, A., 2006, *Virtual migration: The programming of globalization*, Floridi, L., 2016, "Should we be afraid of ai? machines seem to be getting smarter and smarter and much better at human jobs, yet true ai is utterly implausible. Why?", *Aeon*, 9 May 2016. URL = <Floridi 2016 available online> [is this a chapter in a book, or an article?]

Arguedas, A. R., & Simon, F. M. (2023). **Automating democracy: Generative AI, journalism, and the future of democracy**. Balliol Interdisciplinary Institute, University of Oxford.

Asthana, S., & Hazela, A. (2021). Artificial intelligence in education: A study on its applications and challenges. *International Journal of Science and Research*, *11*(4), 1285-1288.

Badoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. AI*, *7*(1), 52–62.

Balibar, E. (2012). Citoyen sujet (et autres essais d'anthropologie philosophique), PUF, trad. it. Cittadinanza, Bollati Boringhieri, Torino 2012.

Balkan Insight (2020). *Berisha claims 'political persecution' as us hits him with sanctions*.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317Bisen, A., Singh, P. K., & Vardia, M. R. (2022). Applications and challenges of artificial intelligence in education. *International Journal of Emerging Technologies in Learning*, *17*(9), 249–260.

Bowman, L. L., Levine, L. E., Waite, B. M., & Dendron, M. (2010). Can students really multitask? An experimental study of instant messaging while reading. *Computers & Education*, *54*(4), 927–931. https://doi.org/10.1016/j.compedu.2009.09.024

Brundage, M. et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Future of Humanity Institute.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In J. Law (Ed). *Power, Action and Belief: A New Sociology of Knowledge* (pp. 196–233). Routledge & Kagan Paul.

Caltagirone, C., (2020). Il desiderio di essere. Per un'"etica del compimento", *Edizioni studium*.

Calvo, P., & Saura García, C. (2024). Against synthetic democracy, available at SSRN: https://doi.org/10.2139/SSRN.4885161

Cameron-Curry, L. (2014). *Educare alla cittadinanza digitale. Un viaggio dall'analogico al digitale e ritorno.* Tangram.

Ceccarini, L. (2021). *The digital citizen(ship). Politics and democracy in the networked society*. Edward Elgar Publishing.Celik, V., Arkın, E., & Sabanci, A. (2023). The use of artificial intelligence technologies in education: A systematic review. Interactive Learning Environments. Advance online publication.

Celot, P., Franceschetti, R., & Salamini, E. (2021). *Educare ai nuovi media. Percorsi di cittadinanza digitale per l'Educazione Civica*, Pearson Academy.Chaiyarak, Y., Paukaew, N., Choosang, P., & Luangphor, S. (2022). Trends and prospects in artificial intelligence education in developing learners' creativity. *Technology, Knowledge and Learning*.

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), 43.

Choi, M., Glassman, M., & Cristol, D. (2017). What it means to be a citizen in the internet age: Development of a reliable and valid digital citizenship scale. *Computers & Education*, 100–112.

Chu, H., & Liu, S. (2023). Can AI tell good stories? Narrative transportation and persuasion with ChatGPT.

Coffé, H., & Bolzendahl, C. (2010). Same game, different rules? Gender differences in political participation. *Sex Roles*, 62, 318–333.

Cogo, G. (2010). *La cittadinanza digitale. Nuove opportunità tra diritti e doveri*. Edizioni della sera, .

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, *32*(3), 444–452.

Costa, P. (2005). Cittadinanza, Laterza.Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Dahal, N. (2023). *Digital citizenship and digital ethics: An educator's perspective*. IGI Global Scientific Publishing.

Danry, V., Leong, J., Pataranutaporn, P., Tandon, P., Liu, Y., Shilkrot, R., Punpongsanon, P., Weissman, T., Maes, P., & Sra, M (2022). AI-generated characters: Putting deepfakes to good use. In S. Barbosa, C. Lampe, C. Appert, & D.A. Shamma (Eds.),.*CHI Conference on Human Factors in Computing Systems* (pp. 1–5). Association for Computing Machinery.

Dasilva, J.P. et al. (2021). Deepfakes on Twitter: which actors control their spread? *Media and Communication* 9.1.

DeHart, J. (Ed.). (2023). *Critical roles of digital citizenship and digital ethics*. IGI Global Scientific Publishing.Dilmurod, N. M., & Fazliddin, M. F. (2022). Integration of artificial intelligence technologies in higher education: Meta-analysis. *Journal of Critical Reviews*, *9*(3), 30–35.

Doshi, A. R., & Hauser, O. P. (2023). Generative artificial intelligence enhances creativity but reduces the diversity of novel content. arXiv preprint arXiv:2312.00506.

Duberry, J. (2022). AI and civic tech: Engaging citizens in decision-making processes but not without risks. *Artificial Intelligence and democracy*, 195–224.

Isin, E., & Ruppert, E. (2015). *Being digital citizens*. Rowman & Littlefield International.

Eco, U., 1984, *Semiotics and the Philosophy of Language*, Indiana University Press.

Ellis, Y., Daniels, W., & Jauregui, A. (2010). The effect of multitasking on the grade performance of business students. *Research in Higher Education Journal*, *8*(1), 1–11.

Emejulu, A., & McGregor, C. (2019). Towards a radical digital citizenship in digital education. *Critical Studies in Education*, *60*(1), 131–147.

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

European Commission (2025). AI act. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai [Accessed March 9th, 2025]

Evidence from hotel reviews. *Journal of Language and Social Psychology*, *43*(1), 63–82.

Exintaris, B., Karunaratne, N., & Yuriev, E. (2023). Metacognition and critical thinking: Using ChatGPT-generated responses as prompts for critique in a problem-solving workshop (SMARTCHEMPer). *Journal of Chemical Education*, *100*(8), 2972–2980.

Federspiel, F., Mitchell, R., Asokan, A., Umana, C., & McCoy, D. (2023). Threats by artificial intelligence to human health and human existence. *BMJ Global Health, 8*(5), e010435. https://doi.org/10.1136/bmjgh-2022-010435

Fernández-Prados, J.S., Lozano-Díaz, A., &Ainz-Galende, A. (2021). Measuring digital citizenship: A comparative analysis. *Informatics,* 8(1). https://doi.org/10.3390/informatics8010018

Floch, J. M. (2001). *Visual Identities*, Continuum.

Floridi, L. (2022). *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*. Raffaello Cortina.Floridi, L., & Sanders, J.W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379. doi:10.1023/B:MIND.0000035461.63578.9d

Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2023). 20160360. doi:10.1098/rsta.2016.0360

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

Franks, B., Bangerter, A., & Bauer, M. W. (2013). Conspiracy theories as quasi-religious mentality: An integrated account from cognitive science, social representations theory, and frame theory. *Frontiers in Psychology, 4*, 424. https://doi.org/10.3389/fpsyg.2013.00424

Gaafar, A. A. (2021). Metaverse in architectural heritage documentation & education. *Advances in Ecological and Environmental Research, 6*(10), 66–86.

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.

Goldsmith, L. P., Rowland-Pomp, M., Hanson, K., Deal, A., Crawshaw, A. F., Hayward, S. E., ... & Hargreaves, S. (2022). Use of social media platforms by migrant and ethnic minority populations during the COVID-19 pandemic: A systematic review. *BMJ Open*, *12*(11), e061896. https://doi.org/10.1136/bmjopen-2022-061896

Gundersen, A. B., van der Linden, S., Piksa, M., Morzy, M., & Piasecki, J. (2023). The role of perceived minority-group status in the conspiracy beliefs of factual majority groups. *Royal Society Open Science*, *10*(3), 221036. https://doi.org/10.1098/rsos.221036

Goodfellow, I., et al. (2014). Generative Adversarial Networks. arXiv preprint arXiv:1406.2661.

Greimas, A. J. (1987). *On Meaning: Selected Writings in Semiotic Theory*. University of Minnesota Press.

Habermas, J. (1991). *Staatsbürgerschaft und nationale Identität. Überlegungen zur europäischen Zukunft*. Erker.

Hajli, N., et al. (2022) Social bots and the spread of disinformation in social media: the challenges of artificial intelligence. *British Journal of Management, 33*, 1238–1253.

Hassoun, A., Beacock, I., Consolvo, S., Goldberg, B., Kelley, P. G., & Russell, D. M. (2023). Practicing information sensibility: How Gen Z engages with online information. arXiv preprint arXiv:2301.07184. https://arxiv.org/abs/2301.07184

Hayes, C. M. (2023). *Generative artificial intelligence and copyright: Both sides of the black box*. Available at SSRN 4517799.

Hembrooke, H., & Gay, G. (2003). The laptop and the lecture: The effects of multitasking in learning environments. *Journal of Computing in Higher Education*, *15*(1), 46–64. https://doi.org/10.1007/BF02940852

Hollanek, T., & Nowaczyk-Basińska, K. (2024) Griefbots, deadbots, postmortem avatars: on responsible applications of generative ai in the digital afterlife industry. *Philosophy & Technology*, *37*(63).

Horovitz, T., & Mayer, R. E. (2021). Learning with human and virtual instructors who display happy or bored emotions in video lectures. *Computers in Human Behavior*, *119*, 106724.

Huang, C., Tu, Y., Han, Z., Jiang, F., Wu, F., & Jiang, Y. (2023). Examining the relationship between peer feedback classified by deep learning and online learning burnout. *Computers & Education*, *207*, 104910.

Huang, R., Liu, D., Tlili, A., Yang, J., Wang, H., et al. (2022). *Handbook on facilitating flexible learning during educational disruption: The Chinese experience in maintaining undisrupted learning in COVID-19 outbreak*. Smart Learning Institute of Beijing Normal University.

Hutchings, K. & Danreuther, R. (Eds.). (1999). Cosmopolitan citizenship. St. Martin's Press.

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* (September), 102714. https://doi.org/10.1016/j.erss.2022.102714.

Karimi, S., Oliveira, M., & Pacheco, D. (2024). Misinformation dissemination: Effects of network density in segregated communities. arXiv preprint arXiv:2411.19866. https://arxiv.org/abs/2411.19866

Kaun, A., Kyriakidou, M., & Uldam, J. (2016). Political agency at the digital crossroads? In A. Kaun, M. Kyriakidou, & J. Uldam (Eds.), *Media and Communication* (pp. 1–7). Cogitatio.

Khurana, M., Chandak, M. B., & Zope, P. A. (2020). *Application of artificial intelligence in education*. 2020 International Conference on Emerging Smart Computing and Informatics (ESCI).

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). Exploring generative artificial intelligence preparedness among university language instructors: A case study. *Computers and Education: Artificial Intelligence*, *5*, 100156.

Kollmann, T., & Kayser, I. (2010). A comprehensive approach to citizen engagement in e-democracy. In *6th International Conference on E-Government–ICEG2006, Academic Conferences and Publishing International, Cape Town,* 54–62.

Kress, G., & van Leeuwen, T. *Reading Images: The Grammar of Visual Design*. Routledge.

Kymlicka, W. & Norman W.J. (Eds.). (2000). Citizenship in diverse societies: Theory and practice, Oxford University Press.

Kymlicka, W. (1995). *Multicultural Citizenship*, Oxford University Press.Lameras, P., & Arnab, S. (2023). Essential features of AI-driven serious games for higher education: A comprehensive review. *British Journal of Educational Technology*. Advance online publication.

Latour, B. (1988). *The Pasteurization of France*. Harvard University Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network theory*. Oxford University Press.

Latour, B. (2017) On actor-network theory: A few clarifications, plus more than a few complications. *Philosophical Literary Journal Logos,* 27(1), 173–197.

Law, J., & Hassard, J. (1999). *Actor-Network Theory and After*. Blackwell.

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

La Torre, M. (2022). *Cittadinanza. Teoria e ideologie*. Carocci.

Leddo, J., & Garg, K. (2021). Comparing the effectiveness of AI-powered educational software to human teachers. *International Journal of Social Science and Economic Research, 6*(3), 953–963.

Lee, H., & Hwang, Y. (2022). Technology-enhanced education through VR-making and metaverse-linking to foster teacher readiness and sustainable learning. *Sustainability*, *14*(8), 4786.

Leiker, D., Finnigan, S., Gyllen, A. R., & Cukurova, M. (2023). Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale. arXivpreprint arXiv:2306.01815.

Leiker, D., Gyllen, A. R., Eldesouky, I., & Cukurova, M. (2023). Generative AI for learning: Investigating the potential of learning videos with synthetic virtual instructors. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 523–529). Springer Nature Switzerland.

Leman, P. J., & Cinnirella, M. (2007). A major event has a major cause: Evidence for the role of heuristics in reasoning about conspiracy theories. *Social Psychological Review, 9*(2), 18–28.

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2012). NASA faked the moon landing—Therefore, (climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science, 24*(5), 622–633. https://doi.org/10.1177/0956797612457686

Leydet, D. (2010). Citizenship, «The Stanford Encyclopedia of Philosophy», a cura di E.N. Zalta, http://plato.stanford.edu/archives/ fall2011/entries/citizenship, 2010.

Liu, L., & Wang, L. (2021). Applications and trends of AIGC technologies in educational field: A review. *Application of Artificial Intelligence*, *2*(1), 012001.

Loeffler, E., & Martin, S. (2015). Citizen Engagement. In *Public Management and Governance*, 3rd ed., Routledge.

Ly-Le, T.-M., & Le, V. T. (2024). Navigating the infodemic: Assessing digital literacy and misinformation vulnerability among senior citizens in Vietnam during COVID-19. *Health & New Media Research*, *6*(1), 25–50. https://doi.org/10.22720/hnmr.2023.00052

Suarez-Villa, L. (2021). The rise of technocapitalism. *Science Studies*, *14*(2), 4–20.

Lythreatis, L., Singh, S.K., & El-Kassar, A. (2022). The digital divide: A review and future research agenda. *Technological Forecasting and Social Change*, 175. https://doi.org/10.1016/j.techfore.2021.121359

Maia, C.H., Ariel, P., & Nunes, S. (2024). Adding human values on the deepfake: Co-designing fact-checking solutions to combat misinformation. *AI and Ethics,* 1–16. https://doi.org/10.1007/s43681-024-00619-y

Mangione, G. (2018). Brief remarks on constitutional court and politics in Italy. *Toruń Polish-Italy Studies*, 21–42. DOI 10.12775/TSP-W.2018.002

Markowitz, D. M., Hancock, J. T., & Bailenson, J. N. (2024). Linguistic markers of inherently false AI communication and intentionally false human communication: [[[[incomplete?]]]]]

Marrone, G., (2022). *Introduction to the semiotics of the text*. De Gruyter Mouton, https://doi.org/10.1515/9783110688986

Martin, B., & Ringham, F. (2000). *Dictionary of Semiotics*. Bloomsbury Academic.

Mazzucato, M. (2019). *Preventing digital feudalism. Project syndicate*. https://www.project-syndicate.org/commentary/platform-economy-digital-feudalism-by-mariana-mazzucato-2019-10

Melo, D.F., & Stockemer, D. (2014) Age and political participation in Germany, France and the UK: A comparative analysis. *Comparative European Politics*, *12*(1), 33–53.

Michalon, B., & Camacho-Zuñiga, C. (2023). ChatGPT, a brand-new tool to strengthen timeless competencies. *Frontiers in Education*, 8, 1251163).

Mittal, U., Sai, S., & Chamola, V. (2024). *A comprehensive review on generative ai for education*. IEEE Access.

Montessori, M. (2017). *Repetition and child development in Montessori Education*. Montessori Academy.

Moore, R. C., & Hancock, J. T. (2022). A digital media literacy intervention for older adults improves resilience to fake news. *Scientific Reports*, *12*, 6008. https://doi.org/10.1038/s41598-022-08437-0Morton, J.L. (2024). On actor-network theory and algorithms: ChatGPT and the new power relationships in the age of AI. *AI and Ethics*, *4*(4), 1071–1084.

Morton, J.L. (2024). On inscription and bias: Data, actor network theory, and the social problems of text-to-image AI models. *AI and Ethics, 1*(1).

Mossberger, K., Tolbert, C.J., &McNeal, R.S. (2007). Digital citizenship: The Internet, society, and participation. MIT Press.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). https://doi.org/10.5334/cstp.583.

Nalivaikė, J., & Miliukaitė, G. (2024). Influence of AI-generated avatars on consumer trust in the brand. *International Scientific Conference Business and Management*. https://doi.org/10.3846/bm.2024.1191.

NATO StratCom Centre of Excellence (2020). Deepfakes and synthetic media: A threat to democratic discourse?

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Comput. Educ. Artif. Intell.* 2, 100020. doi: 10.1016/j.caeai.2021.100020

Parry, G., Moyser, G., & Day, N. (1992). *Political Participation and Democracy in Britain*. Cambridge University Press.

Participation. *ICEGOV*. https://doi.org/10.1145/3680127.3680194

Pascucci, G. (2021), *La cittadinanza digitale. Competenze, diritti e regole per vivere in rete*. Il Mulino.

Patel, R. (2025). Strategies for Including Public Voices in AI. *AI for Digital Democracy: How to navigate opportunities and threats*, Webinar https://www.peoplepowered.org/events-content/ai-and-digital-democracy-2025

Pinto, F., & Macadar, M. A. (2024). Using generative ai chatbot for enhancing people's digital [[[incompleet?]]]]

Polidoro, P., (2017). Spectator in fabula: A model of the interpretative cooperation in visual texts. In T. Thellefsen & B. Sørensen, B. (Eds.). *Umberto Eco in His Own Words*. De Gruyter.

Polidoro, P. (2018). *Post-Truth and Fake News. Preliminary Considerations*. Versus 2/2018, https://www.rivisteweb.it/issn/0393-8255/issue/7537.

Politico EU (2018). "How Albania's politicians courted Donald Trump."

Pollicino, O., Bertolini, E., & Lubello, V. (Eds.). (2013). Internet: Regole e tutela dei diritti fondamentali. Aracne.

Popenici, S., & Kerr, S. (2024). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning, 12*(1), 22

Poulsen, S.V. (2021). Face off – A semiotic technology study of software for making deepfakes. *Sign Systems Studies*, *49*(3-4): 489–508. DOI:10.12697/SSS.2021.49.3-4.12

Quintelier, E. (2007). Differences in political participation between young and old people. *Contemporary Politics*, *13*(2), 165–189. doi: 10.1080/13569770701562658

Rawal, M. (2020). National Education Policy 2019: For the transferable job parents which affect their children's education. *Educational Resurgence Journal*, 2(3).

Ren, J., Xu, H., He, P., Cui, Y., Zeng, S., Zhang, J., ... & Tang, J. (2024). Copyright protection in generative ai: A technical perspective. arXiv preprint arXiv:2402.02333.

Reuters. (2024, June 3). *European election: How the EU says Russia is spreading disinformation*. Reuters. https://www.reuters.com/world/europe/european-election-how-eu-says-russia-is-spreading-disinformation-2024-06-03/

Riva, F. (2020). *Etica e cittadinanza*. Edizioni Lavoro.

Robb, M. B. (2020). *Teens and the news: The influencers, celebrities, and platforms they say matter most*. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/uploads/research/2020_teens_and_the_news.pdf

Robinson, D., Delany, J., Sudgen, H. (2024). Beyond science: Exploring the value of co-created citizen science for diverse community groups. *Citizen Science: Theory and Practice*, *9*(1), https://doi.org/10.5334/cstp.682

Rodriguez, R., & K, H.[[[[incomplete?] (2023). The future of education: Exploring AI avatars in higher learning. *Qeios*. https://doi.org/10.32388/80z989.

Ronchi, A. M. (2019). Being Human in the Digital Age: E-Democracy. In A.M. Ronchi (Ed.), *E-Democracy: Toward a New Model of (Inter)Active Society* (pp. 1–4). Springer International Publishing. https://doi.org/10.1007/978-3-030-01596-1_1.

Rodotà, S. (2014). *Il mondo nella rete. Quali i diritti, quali i vincoli*. Laterza.

Rovetta, F. & Valentini, D. (2025) Grief and virtual reality: Continuing bonds with virtual avatars. *Phenomenology and the Cognitive Sciences*.

Sandfort, J. & Quick, K. S. (2015). Building deliberative capacity to create public value. In J.M. Bryson, B.C. Crosby, &L. Bloomberg (Eds.), *Public Value and Public Administration* (pp. 39–52). Georgetown University Press.

Schlozman, K. L., Burns, N., Verba, S., & Donahue, J. (1995). Gender and citizen participation: Is there a different voice? *American Journal of Political Science*, *39*(2), 267–293.

Schneider, S., Krieglstein, F., Beege, M., & Rey, G. D. (2022). The impact of video lecturers' nonverbal communication on learning – an experiment on gestures and facial expressions of pedagogical agents. *Computers & Education*, *176*, Article 104350.

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, *49*(1), 1–39.

Shen, J., & Su, Y. (2021). The role of artificial intelligence in helping teachers improve education and teaching. *Frontiers in Psychology*, *12*, 645910.

Somoray K, & Miller DJ (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

Sondermann, C., Huff, M., & Merkt, M. (2024). Distracted by a talking head? An eye tracking study on the effects of instructor presence in learning videos with animated graphic slides. *Learning and Instruction*, *91*, Article 101878.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778. https://doi.org/10.1126/science.1207745

Su, J., & Yang, W. (2024). Powerful or mediocre? Kindergarten teachers' perspectives on using ChatGPT in early childhood education. *Interactive Learning Environments*, *32*(10), 6496–6508.

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022). Investigating explainability of generative AI for code through scenario-based design. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (pp. 212–228).

Tack, A., & Piech, C. (2022). The AI Teacher test: Measuring the pedagogical ability of Blender and GPT-3 in educational dialogues. arXiv preprint arXiv:2205.07540.

Tang, Z., Goh, D. H.-L., Lee, C. S., & Yang, Y. (2024). Understanding seniors' strategies for identifying deepfakes. In *HCI International 2024 Posters* (pp. 236–244). Springer. https://doi.org/10.1007/978-3-031-61947-2_26

Tauginienė, L., Butkevičienė, E., Vohland, K., Heinisch, B., Daskolia, M., Suškevičs, M., Portela, M., Bálint B., & Prūse, B. (2020). Citizen science in the social sciences and humanities: the power of interdisciplinarity. *Palgrave Commun, 6*, 89. https://doi.org/10.1057/s41599-020-0471-y

Taylor, D.,Turner, B., & Hamilton, P. (Eds.). (1994). *Citizenship: Critical Concepts*. Routledge, .

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian Paternalism. *American Economic Review, 93*(2), 175–179. doi: 10.1257/000282803321947001.

The New York Times (2017). "Trump's Balkan Business Connections."

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D.C.M., Ezer, D., van der Hart, F.C., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., de Wever, W., … & Clopath, C. (2020). AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, *11*(1), 2468. https://doi.org/10.1038/s41467-020-15871-z

Toolan, N. (2018). *3 ways the fourth Industrial Revolution is disrupting the law*. World Economic Forum.

Tsai, L. L., Pentland, A., Braley, A., Chen, N., Enríquez, J. R., & Reuel, A. (2024). Generative AI for pro-democracy platforms. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.5aaf489a

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

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

United Nations (b. d.). The 17 goals. https://sdgs.un.org/goals [Accessed March 9th, 2025]

Unver, M. B. (2024). AI governance: Compromising democracy or democratising AI? *Proceedings of the TPRC2024 The Research Conference on Communications, Information and Internet Policy*. https://ssrn.com/abstract=4913658

Valera, L., & Castilla, J.C. (Eds). (2020). *Global changes. Ethics, politics and environment in the contemporary technological world*. Springer.

Varoufakis, Y. (2023) *Technofeudalism: What killed capitalism*. Random House.

Vartiainen, H., & Tedre, M. (2023). Using artificial intelligence in craft education: Crafting with text-to-image generative models. *Digital Creativity*, *34*(1), 1–21.

Vivion, M., Reid, V., Dubé, E., Coutant, A., Benoit, A., & Tourigny, A. (2024). How older adults manage misinformation and information overload: A qualitative study. BMC Public Health, 24, 871. https://doi.org/10.1186/s12889-024-18335-x

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Rerelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (2021). *The Science of Citizen Science*, Springer International Publishing

Von Nostitz, F., Neihouser, M., Sandri, G., & Haute, T. (2024). Patterns and factors of political disconnection on social media: A cross-platform comparison. *Media and Communication, 12*, Article 8544. https://doi.org/10.17645/mac.8544

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 39–52. http://doi.org/10.22215/timreview/1282

Williams, S., Beery, S., Conley, C., Evans, M. L., Garces, S., Gordon, E., Jacob, N., & Medina, E. (2024). *People-powered Gen AI: Collaborating with generative ai for civic engagement. An MIT exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.f78710e6.

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. arXiv preprint arXiv:1902.11097.

Wineburg, S., & McGrew, S. (2016). Evaluating information: The cornerstone of civic online reasoning. Stanford History Education Group. https://purl.stanford.edu/fv751yt5934

Winner, L. (1993). Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology. *Science, Technology, & Human Values, 18*(3), 362–378.

Wörsdörfer, M. (2024). Datafeudalism: The domination of modern societies by tech companies': A commentary. *Philosophy and Technology*, *37*(3), 1–5 https://doi.org/10.1007/S13347-024-00781-5

Xing, Z., Yuan, X., & Vizer, L. (2020). The age-related differences in web information search process. arXiv preprint arXiv:2010.13352. https://arxiv.org/abs/2010.13352

Xu, T., Gao, Q., Ge, X., & Lu, J. (2024). The relationship between social media and professional learning from the perspective of pre-service teachers: A survey. *Education and Information Technologies*, *29*(2), 2067–2092.

Yazar, T., & Arifoglu, G. (2012). A research of audio visual educational aids on the creativity levels of 4-14 year old children as a process in primary education. *Procedia-Social and Behavioral Sciences*, *51*, 301–306.

Yu, X., Liu, L., Hu, X., Keung, J. W., Liu, J., & Xia, X. (2024). Fight fire with fire: How much can we trust ChatGPT on source code-related tasks?. *IEEE Transactions on Software Engineering*.

Zhang, A., Walker, O., Nguyen, K., Dai, J., Chen, A., & Lee, M.K. (2023). Deliberating with AI: Improving decision-making for the future through participatory AI design and stakeholder deliberation. *Proceedings of the ACM on Human-Computer Interaction, 7*, 1–32. https://doi.org/10.1145/3579601

Zhang, C., Zhang, C., Zheng, S., Zhang, M., Qamar, M., Bae, S. H., & Kweon, I. S. (2023). A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. arXiv preprint arXiv:2303.13336.

Zheng, W., Lu, S., Cai, Z., Wang, R., Wang, L., & Yin, L. (2023). PAL-BERT: an improved question answering model. *Computer Modeling in Engineering & Sciences*, *10*.

Zhou, J., Xiang, H., & Xie, B. (2023). Better safe than sorry: A study on older adults' credibility judgments and spreading of health misinformation. Universal Access in the Information Society, 22, 957–966. https://doi.org/10.1007/s10209-022-00899-3

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668.

Ziccardi, G. (2012). *Resistance, liberation technology and human rights in the digital age*. Springer, .Zuboff, S (2019), *The age of surveillance capitalism: The fight for the future at the new frontier of power*. Profile Books.