SOLARIS

Strengthening democratic engagement through value-based generative adversarial networks

D2.2 SOLARIS
Socio-political and psychological attitude toward GANs

Lead Author: Piero Polidoro (LUMSA)
With contributions from UVA, CINI, DEXAI, UM, ECSA
Reviewers: BMB-EXE

Deliverable nature:	R-Document
Dissemination level: (Confidentiality)	PU - Public
Delivery date:	30 November 2023
Version:	Final
Total number of pages:	49
Keywords:	Deep Fakes, Generative AI, ANT, AI for good, Psychometric Scale, Visual Semiotics

Executive summary

A preliminary part is devoted to summarising the results of the previous report, D.2.1.
Chapter 1 presents various up-to-date types of neural networks and investigates the ethical problems associated with AI.
Chapter 2 of the report presents the project's theoretical framework. Our approach to understanding how social media users perceive and trust AI-generated content is introduced. We use the philosophy of technology and information, as well as science and technology studies, to build on the Actor-Network Theory approach of Bruno Latour (ANT). Based on this, we describe the social network in which a user is embedded and reframe the question of 'trust' in AI-generated content from a system-level perspective. This deliverable focuses specifically on a semiotic analysis of deepfakes. We assume that deepfake images or videos are never viewed in isolation. They are enjoyed through specific social discourse and embedded in communicative flows. A deepfake is influenced by its immediate context and can be part of a chain of cause-and-effect or before-and-after relationships with other textual objects. The internal structure of an image is organised on several semiotic levels. For example, a deepfake always combines visual and verbal language. Deepfake videos, on the other hand, combine visual, verbal, and sound languages.
In Chapter 3, we create a valid scale that can measure the different features that affect the perceived trustworthiness of GANs. The scale focuses on content-related aspects that determine how likely it is for GAN-generated content to be considered trustworthy, as opposed to individual characteristics that can be measured using existing scales. We adopt the Open Science approach to enhance research data management through open and collaborative methods of generating and sharing knowledge and data. The new EU approach will affect research institutions and science practices by introducing new funding, evaluation, and reward systems for researchers. Within this framework, the SOLARIS data management plan adopts the FAIR approach (findable, accessible, interoperable, and reusable).
Chapter 4 presents the most promising perspectives for the positive use of generative AI, including some examples and analyses.

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**	D.2.2	**Title**	Socio-political and psychological attitude
Work package	**Number**	2	**Title**	Definition& Mapping of GANs in social media
Task	**Number**	2.3	**Title**	Socio-political and psychological characteristics of GANs

Date of delivery	**Contractual**	M10	**Actual**	M10
Status	version 1		☒Final version	
Nature	☒Report ☐Demonstrator ☐Other ☐ORDP (Open Research Data Pilot)			
Dissemination level	☒Public ☐Confidential			

Authors (partners)	UVA, CINI, LUMSA, DEXAI, UM, ECSA			
Responsible author	**Name**	Piero Polidoro		
	Partner	LUMSA	**E-mail**	p.polidoro@lumsa.it

Summary (for dissemination)	
Keywords	Deep Fakes, Generative AI, ANT, AI for good, Psychometric Scale, Visual Semiotics

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

Acknowledgement

Disclaimer

Table of contents

List of Figures

List of tables

Abbreviations and acronyms

Abbreviation	Meaning
AI4SG	Artificial intelligence for social good
ANT	Actor Network Theory
CC BY 4.0	Creative Commons Attribution 4.0 International
Celeb-DF, DFD, DFDC	Publicly available datasets developed for deepfake detection
CVI	Content Validity Index
CVR	Content Validity Ratio
EC	European Commission
FAIR data	Data which meet principles of findability, accessibility, interoperability, and reusability
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
NLG	Natural Language Generation
PIG	Psychometric Item Generation
VAE	Variational Autoencoders

1 Generative AI and the problem of deep fakes

The previous report, D.2.1., delivered on 30 September was the first attempt to define the Adversarial Generative Network types developed from 2017 to 2023. This report was an attempt to do the first mapping of Generative AI addressing definitions, models, and types of contents exploring negative consequences and challenges to our social and democratic systems at national and international level. Specific attention is given to the policies and regulatory mechanisms that should contrast the eventual neatives effects of generated AI and deep fakes, in Europe and elsewhere.

A link was drawn between AI-generated content, deepfakes and Generative Adversary Networks (GANs). Generative adversarial networks (GANs) are a class of AI models able to create media content – audio and video – designed to minimize the distance between reality and antificially generated contents.

How did we get from AI-generated content to deepfakes: Generative AI is adept at producing new data instances from existing ones, such as realistic avatars. In contrast, Discriminative models differentiate between data. Deepfakes, powered by Generative Adversarial Networks (GANs), create content that often appears authentic and has varied applications.

Although different promising areas of application of GANs – e.g. audio-graphic productions, human-computer interactions, satire, artistic, creative expression – could also mean a revolution with far-reaching consequences in many societal sectors (education, health, legal, economic, and climate crisis), their current and foreseen misleading uses are just as numerous and worrying. The main concern concerns the so-called “deepfakes”, fake images or videos simulating real events with extreme precision. If trained on a face, GANs can make it move and speak hyper-realistically. This technology poses an urgent political threat since GANs could be – and have already been – used to spread fake news and disinformation.

The negative policy implications of these technologies were explored, particularly concerning their abuse or misuse and the biases they can foster. There is a whole working area devoted to the study of the risks that Generative AI poses at the level of international relations since GANs and AI have also started to create internal and external problems being utilized in electoral campaigns, damaging the opponent’s reputation or exacerbating tensions where they already exist. This issue raises an urgent challenge to democratic governance and regulation: to improve GANs accountability, transparency, and trustworthiness.

SOLARIS reacts to these challenges in two ways. On the one hand, we analyzed political risks associated with these technologies to understand and prevent negative implications for EU democracies. There is a negative impact of Deepfakes on democracy. Generative AI can produce misleading media that pose significant threats to democracy, misrepresenting public figures and potentially altering political sentiments. Furthermore, AI-generated media has the potential to intimidate individuals, exploit credibility, manipulate geographies, and influence political discourse.

Threats can also come from natural language generation (NLG) tools. Tools such as ChatGPT can in fact produce persuasive texts, thus enhancing disinformation campaigns. AI applied to large language models is a field of prolific development highly influenced by financial investment and it is needs further regulatory containment.

The widespread use of AI-generated content can erode trust in genuine information sources. As distinguishing between real and fake AI content becomes more challenging, individuals may rely more on personal beliefs than factual information, leading to barriers in communication, misconceptions about AI's capabilities, and an aversion to adopting AI for beneficial purposes. As a result, SOLARIS is already envisaging specific regulatory innovations to detect and mitigate

Generative Adversarial Networks for which we still do not know the involved risks, neither the level of risks.

On the other hand, we assess the opportunities raised by GANs for reinvigorating the democratic engagement of citizens. We will co-create, involving citizen science, value-based GANs contents to enhance democratic engagement, associated today with a myriad of issues for which information or misinformation is crucial, human migration, the climate crisis, to enhance active and inclusive digital citizenship ultimately. SOLARIS will co-create, involving citizen science, value-based GANs contents to enhance democratic engagement.

To achieve these results an exploration of the link between deepfakes and fake news from a semiotic perspective was presented in D.2.1. the first mapping of AI generated content with eventual negative impact models were constructed to articulate different types of lies and fakes, discussed concerning deepfake images. In the current post-truth era, emotions often trump objective facts in shaping public opinions. It's vital to differentiate between unintentional falsehoods (misinformation) and deliberate deception (disinformation). The rise of misleading deepfakes has led to the emergence of roles like debunkers and fact-checkers.

Due to the complex and multi-dimensional nature of the issue, a cross-legal approach was adopted in D.2.1, considering its impact on various legal areas. In fact, the current EU regulatory framework faces challenges in addressing the broad implications of Generative AI. The EU has proposed categorising AI risks to balance innovation and public trust. Ongoing debates revolve around the risk categorization of Generative AI, with recommendations for the future leaning towards the adoption of a comprehensive ethical and moral directive for AI and possibly setting up an overarching EU AI agency. An overview of the global situation was provided: The USA lacks a federal AI regulation, fearing it might hamper their lead in a burgeoning market, particularly against China. However, states such as California have adopted acts like the Artificial Intelligence Accountability Act of 2022. Industry figures, such as Sam Altman from OpenAI, are vocal about the need for regulation. China is proactively examining AI regulation, transitioning from specific rules to a draft general AI law. Their main guiding tool is the 2017 New Generation AI Development Plan, acknowledging the benefits of early regulation. Despite aggressive AI deployment, India doesn't have a comprehensive regulatory system. The NITI Aayog agency suggests establishing an AI regulatory structure, triggering extensive discussions. Post-Brexit, the UK's approach to AI regulation veers from the EU's. The UK's 2023 White Paper stresses trust in AI, suggesting a non-statutory, sector-specific approach, encompassing principles like safety, transparency, and accountability.

It's important to include public governance methods in the remedies we consider addressing various threats, such as those to democracy and reputation. However, the EU is struggling to implement AI transparency due to technical challenges. In this regard, the European Parliament is calling for transparency in Generative AI, particularly in terms of content disclosure and prevention of illegal content. As the process of establishing rules can be complicated and time-consuming, it's recommended to use AI Pacts and voluntary codes of conduct to tackle the potential issues related to Generative AI.

In the last chapter of D.2.1, SOLARIS provided examples of Generative AI in media landscapes, e.g. Bulgaria was taken as an example. The report highlighted Bulgaria as the least media literate country in the EU, and therefore highly susceptible to misinformation. Due to its close ties with Russia and a media scene largely controlled by a few entities, the country is vulnerable to disinformation. Although deepfakes are not prevalent in Bulgaria, it could be due to the resources required to create them or the effectiveness of more basic manipulations. Platforms like Facebook have implemented measures against deepfakes since 2020, which has lessened their production motivation. Fortunately, existing manipulation tools are relatively easy to detect, but the future is uncertain as technology continues to advance.

1.1. Technical Analysis

The field of deepfake technology is rapidly advancing, with new techniques and tools being developed to create more realistic and convincing video and audio forgeries. This section will recap some of the latest techniques for creating deepfake and the technical challenge associated with each, already presented in D2.1.

1. GANs are a common technique for creating deepfakes. GANs involve two neural networks: a generator network that produces the fake video or the audio, and a discriminator network that identifies which samples are real and which are fake. By training the generator to create more convincing samples that can fool the discriminator, GANs can produce multimedia content which seems realistic. Thanks to their architectural scheme GANs can manipulate many different types of multimedia content as text, but the most used one is the Transformer Model. However, there are several technical challenges associated with GANs, including:

- Mode collapse

2. Speech synthesis techniques, which can be used to create convincing audio forgeries. Speech synthesis techniques involve adversarial training of neural networks to create speech that sounds person's voice which is used for analyze the voice's samples. Some of the technical challenges associated with the speech synthesis techniques include:

- Prosody and intonation

3. Face Swap involves transferring the movements and facial expressions of one person to another in a video. These techniques can be used to create convincing deepfakes that show a dialogue or an action that the person did not say or did not do.

4. Face Reenactment involves the manipulation of a user's facial expression in a deepfake image, e.g. to add or remove a smile.

Some of the technical challenges associated with motion transfer techniques include:

- Cross-domain transfer

5. Few-shot Learning is a technique that allows the model to learn from a limited set of examples, making it easier to create a deepfake with less training data. The main challenge of this technique is limited data.

6. There are other different architectures of AI models that allow the creation of deepfakes, most commonly "Variational Autoencoders (VAE)" and "Diffusion Model".

VAEs are a type of neural network that consist of two parts: an encoder and a decoder. The encoder and the decoder are two specific types of neural network. The encoder, in particular, can learn to compress data, such as images, audio or text; while the decoder generates new data starting from data given by the encoder. The VAEs are trained to compress input data into a lower-dimensional representation, called a latent space in order to reconstruct the data from this representation.

The technical challenges of this technique include:

- Data augmentation

Diffusion model

Compared to GANs, the diffusion models involve the creation of an image from random noise. The training process consists in the creation of the desired image, transforming the noise into the image. At this stage, the model can learn the structure of the data and reproduce the generation. During the training, the model tries to minimise the error between the image generated by the noise and the actual sample. An example of these models are DALL-E or Midjourney. GANs are widely used in comparison to the diffusion models due to their capacity of learning the characteristic details of the images to be generated.

1.2. Ethical Analysis

Deepfakes are a manifestation of technology with profound social ramifications. They emerged primarily to emulate individuals for deceptive purposes, often with malign intent. With the advent of Generative Adversarial Networks (GANs), the first technology capable of creating deepfakes, predominant uses have alarmingly involved generating non-consensual pornographic material, fabricating politically damaging videos, and other uses with significant ethical concerns. Ethically, deepfakes pose numerous challenges and dilemmas. The SOLARIS project is centered on addressing these ethical and social dangers, striving to harness deepfake technology for socially constructive purposes while avoiding ethical pitfalls. Consequently, a thorough examination of the ethical dimensions is critical for SOLARIS, and the entire initiative is committed to mitigating these issues. In this document, we delve into the psychological and political aspects of deepfakes from a theoretical standpoint and discuss the ethical implications that emerge from the material presented herein.

The ethical dimensions of deepfakes:

The capacity of deepfakes to deceive is at the root of their ethical concern. These arise from the deepfakes' potential to undermine the autonomy of individuals to discern truth from falsehood. When people are presented with manipulated media that is indistinguishable from authentic content, their ability to make informed judgments based on truthful information is compromised. This has serious implications, as it can weaken the trust that is essential for personal relationships and for the functioning of a democratic society. Deepfakes challenge our understanding of truth and authenticity. When a deepfake video is almost indistinguishable from a real one, it becomes difficult to uphold the notion of an objective reality. This has serious implications for fields like journalism and law enforcement, where visual evidence plays a critical role. Fabricated content in these areas can have far-reaching consequences, including the corruption of the historical record, the miscarriage of justice, and the undermining of public trust in essential institutions. The issue of consent is also paramount when it comes to deepfakes. Using someone's likeness without their agreement, particularly for harmful purposes, violates personal rights and dignity. This is seen most clearly in cases where deepfakes are used to produce pornographic material or to slander individuals, actions that can cause irreparable damage to people's lives and careers. The potential use of deepfakes in international relations adds another layer of complexity to the ethical debate. They could be used to create false evidence, to mislead the public or international community, and potentially to provoke conflicts or exacerbate tensions between nations. The question of accountability is also crucial. There needs to be a clear framework for holding individuals or groups responsible for creating or spreading deceptive deepfake content. However, the anonymous nature of the internet makes it difficult to trace the origins of such content, allowing those who wish to deceive to do so with less risk of being caught. The SOLARIS project aims to address these ethical challenges by finding ways for deepfakes to be used for good, while also creating safeguards against their misuse. The goal is to redirect the technology away from deception and towards applications that benefit society. However, achieving this is complex, requiring careful ethical consideration and robust regulatory frameworks to prevent abuse. It is important to note that the possible ethical issues related to the implementation of SOLARIS project are already addressed in D1.2, and therefore will not be reported here.
The analysis explores the various ethical issues raised by the deepfake phenomenon. This may include discussions of:

Deception is the primary and most evident concern regarding the production of deepfake content. Initial discussions about the deceptive nature of technologies were concentrated on the potential for social robots to mislead users, as explored by Sharkey & Sharkey (2021) and Sparrow & Sparrow (2006). Although current social robots are not mistakable for humans, the issue of deception was already a point of contention. Deepfakes, however, elevate the problem of deception to an unprecedented level. The SOLARIS project seeks to discern the conditions under which people deem generated content to be trustworthy, working towards developing a psychometric scale that assesses perceived trustworthiness.

Societal Oversight: Another issue exacerbated by deepfakes is the need for societal oversight of their effects. Deepfakes pose a significant challenge to detection, both by human evaluators and AI systems designed to differentiate authentic content from falsifications. Although certain technical countermeasures, such as Intel's deepfake detector, appear to be effective, they may not suffice given the overwhelming volume of content produced. Currently, deepfake detection technologies can manage a convenient quantity of content and are employed by authoritative entities such as news platforms. However, such technical solutions are not as effective on social media and the web at large. SOLARIS aims to bridge this gap by proposing a socio-technical approach to deepfakes, encompassing both technical strategies and sociological tactics to mitigate associated risks.

Privacy: One of the most significant ethical concerns associated with deepfakes is the potential impact on an individual's privacy. Deepfakes can be used to create images or videos of the individuals in intimate or sensitive situations, such as engaging in sexual acts or making inflammatory statements. These images or videos can then be shared on social media or other platforms, potentially causing significant harm to the individual's reputation and personal life.
Furthermore, deepfakes can be used to create fake pornography or revenge porn, which is a form of sexual harassment and a violation of privacy. Even if the individual depicted in the deepfake is not real, the images or videos can still be used to harass or intimidate individuals who resemble the person in the deepfake. To address these privacy concerns, there have been calls for legislation to regulate deepfakes and prevent their malicious use. The transparency rule has proofed to be not enough to contrast pernitious consequences.

Consent: Related to the issue of privacy there is the importance of obtaining consent from individuals before creating or sharing deepfakes that depict them. In many cases, individuals who are depicted in deepfakes have not given their consent to be depicted in such a manner. This raises significant ethical concerns, as it represents a violation of the individual's autonomy and right to control their own image. In order to address this issue, some experts have called for clear guidelines and standards regarding the use of the deepfakes, as well as increased education and awareness about the potential harm caused by their misuse.

Trust: Another critical ethical issue associated with deepfakes is the potential impact on trust in information and media. As deepfakes become more advanced and realistic, it may become increasingly difficult to distinguish between real and fake media. This has the potential to erode trust in media and information sources, as well as in the authenticity of public figures and institutions.
Furthermore, deepfakes can be used to spread false or misleading information, such as in the form of political propaganda or disinformation campaigns. This represents a significant threat to democratic institutions and the ability of individuals to make informed decisions based on accurate information. To address these trust-related concerns, it is essential to develop methods for detecting and verifying the authenticity of media and promote media literacy and critical thinking skills among the public.

Manipulation of public opinion: Perhaps the most significant ethical concern associated with deepfakes is the potential for manipulation and harm. Deepfakes can be used to manipulate public

opinion, sow discord, and undermine trust in democratic institutions For example, deepfakes can be used to create false or misleading videos of political candidates, which could influence the outcome of an election. Furthermore, deepfakes can be used to manipulate individuals on a personal level. For example, a deepfake of a family member or friend could be used to extract sensitive information or manipulate an individual's emotions. To address these manipulation-related concerns, it is important to develop ways to detect and prevent the malicious use of deepfakes.

Accountability: Another pressing concern centers on the accountability of individuals for their actions in the context of deepfake technology. The capability of deepfakes to convincingly replicate human appearance and behavior presents two inversely related challenges to accountability. On one hand, individuals, particularly those in the public eye such as politicians and celebrities, may be wrongly accused of actions they did not commit—a phenomenon with significant repercussions, as detailed in the cases outlined in section D2.1 of our report. Conversely, the same technology affords individuals the opportunity to deny responsibility for their actual deeds, attributing them to the creation of deepfakes.
The dilemma is exacerbated by the limitations of current detection technologies. If content is identified as a deepfake through technical measures, it can exonerate a person wrongly accused. However, no system can yet guarantee with absolute certainty the authenticity of digital content. The concern arises particularly with high-quality deepfakes that evade current detection capabilities. Consequently, it becomes possible for anyone to dispute the veracity of content in which they are depicted, thus posing a significant challenge to the assignment of responsibility. The difficulty lies in establishing a reliable method to discern genuine actions from sophisticated forgeries, an issue that remains unresolved and deeply problematic

2 Theoretical Framework

In this section, we present the theoretical framework of the project. We first introduce our distinct entry point into the question of how social media users come to believe, or trust, AI-generated content. We approach this question from the perspective of philosophy of technology and information and of science and technology studies, and we build on the 'Actor Network Theory' (ANT) approach of Bruno Latour. On this basis, we describe the network of social relations in which a user is embedded, and we reframe the question of 'trust' in AI-generated content from a network or system-level perspective.

2.1. A Systemic Level Approach to AI Generated Contents

On social media the circulation of AI-generated content is increasing at rapid speed (see also D2.1 for a relevant analysis). Motivated by the potential threats to democratic processes and geopolitical instability and the potential erosion of individual trust described in D2.1, SOLARIS sets out to ask how users come to trust AI-generated content.
SOLARIS' approach is premised on the idea that trust in AI-generated content is *not* solely determined by the technical quality of audio-visual content produced. Instead, any AI-generated content is part of a broad network or system, that includes developers, social media infrastructures, individual users and their environment, institutions of various kind, etc.

In the literature in the philosophy of technology and philosophy of information, there are accounts that attempt to describe and analyze systems that are hybrid, namely that involve both humans and artefacts – these are called socio-technical systems (Abbas, 2023)(, or in the account developed by Bruno Latour, we deal with a network of different actors (1987). Simply put, ANT sees every reality as a 'flat' network of actors, or better said actants, that can be human, artificial/synthetic, natural objects, social institutions, values, or others. It is 'flat' because it does not impose a pre-determined hierarchy between these actors. With ANT, we can describe *any system*, and in particular, SOLARIS is interested in describing systems in which AI-generated contents, such as deepfakes, are produced and shared on social media. In section 2.2.1, we describe the network of actor at work in the case of social media and AI-generated content.

2.2 The Actor Network

At the moment of viewing deepfake content in an online environment, a number of social actors (actors can be human, non-human, artefacts, organisations, individuals) converge and interact with one another in a flat non-hierarchical network of relationships. In order to assess the influence of AI-generated content, these social actors and the nature of their interactions must be analysed through a number of multi-disciplinary approaches. This network of actors is expansive and complex with numerous social actors involved and all linked together by intricate and ever-changing relationships. As such, the diagrams presented in Figure 1 and 2 are simplified versions in order to identify the key social actors at play, to elaborate on their different characteristics, and to illustrate how these actors are linked within the network. Figure 1 provides a general approximation of the different groupings of social actors and how these groupings are generally understood to relate to one another. These include those social actors involved in the development and distribution of a generative AI program, the creation of deepfake content using this program, the target themselves, the circulation of this content in online spaces, the user reception of the content, the various policy and legislative interventions, and the broader public discourse surrounding the deepfake content. Each of these groupings can be further unpacked to reveal a multitude of social actors at play, as shown in Figure

2. The following section will elaborate on these groupings and the characteristics of the social actors in operation.

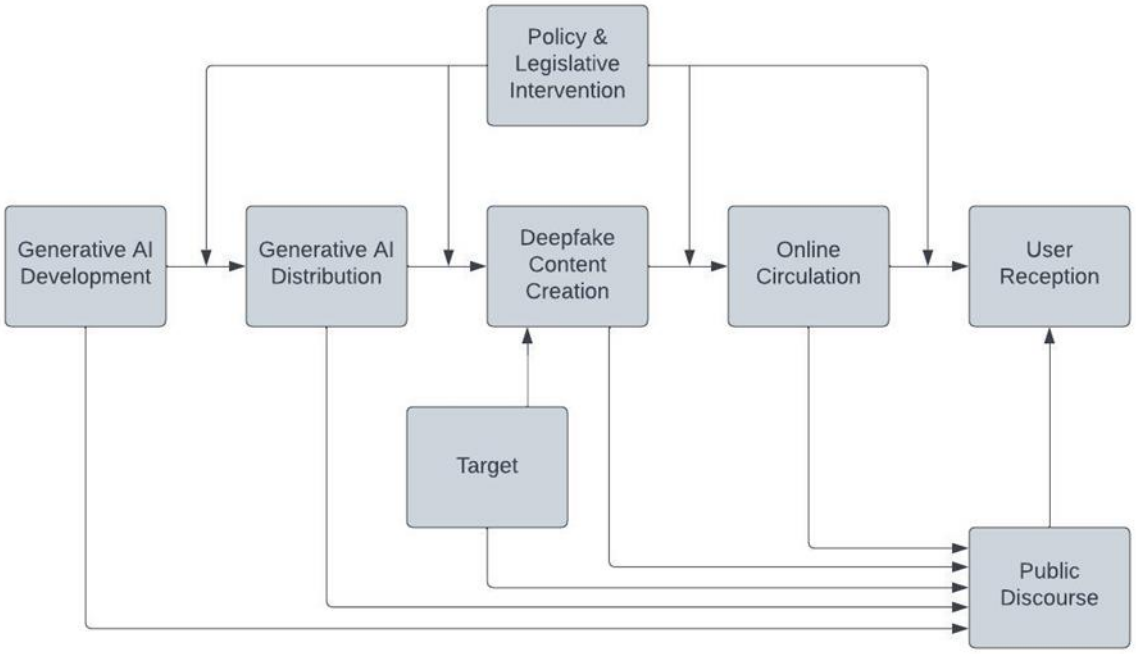

Figure 1. Approximation of groups of social actors involved in the production, circulation and reception of deepfake content.

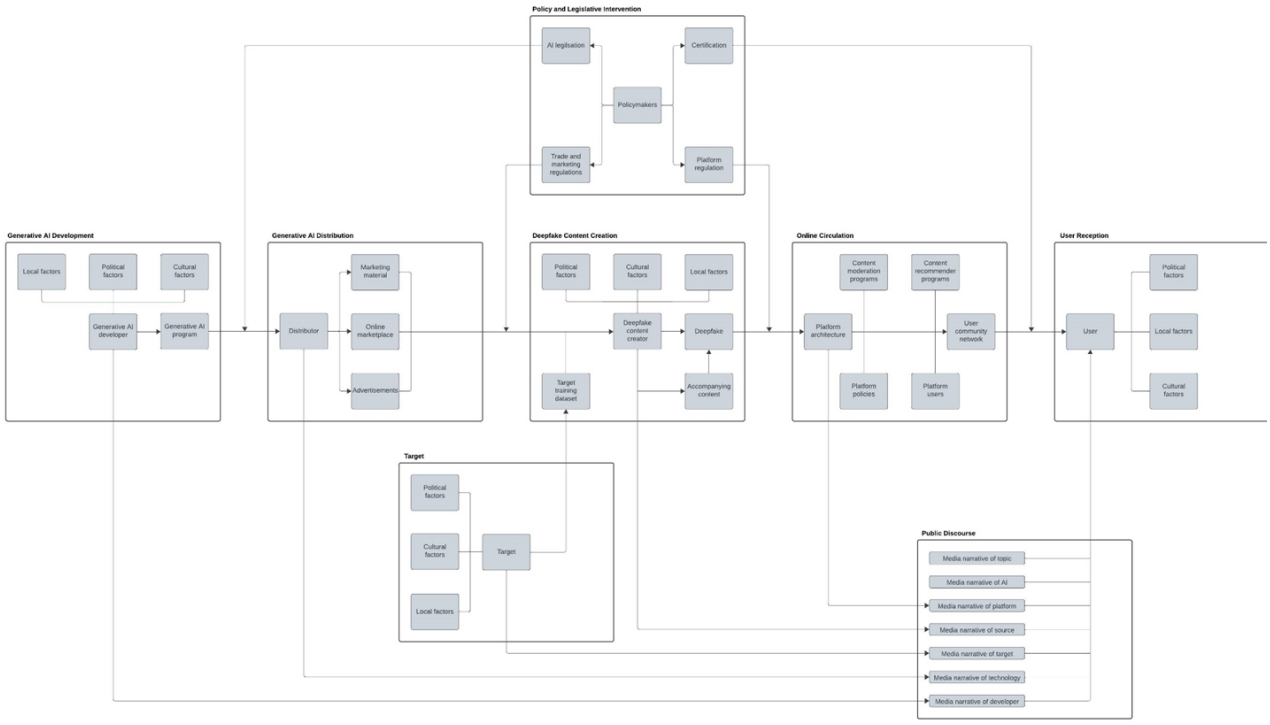

Figure 2. Expanded view of social actors involved in deepfake actor-network.

Generative AI development

This grouping accounts for those social actors involved in the development and production of the generative AI program itself, primarily the characteristics and motivations of the generative AI developer (e.g., private company, government institution, research centre) and those factors within the developer's social environment that impact their actions and practices (e.g., government regulation, competition).

- *Generative AI developer:* there are many different types of generative AI developer (e.g., government institution, private company, public service provider, academic research center, independent programmer) with different associated characteristics that may influence or alter their role as a social actor within the actor-network. These characteristics include the developer's intention for producing a generative AI program (e.g., profitable product, public service, industry innovation), their accessibility to resources (e.g., funding, expertise, ethical training), and their adherence to value-based industry standards (e.g., OECD AI Principles) and regulations (e.g., GDPR, EU AI Act).
- *Developer social environment:* the actions of generative AI developers are also determined, to some extent, by various political, cultural and social factors at play in their local and general environment. As developers may be understood as national institutions, multi-national corporations, or independent citizens, these factors may differ significantly.
    - *Political factors*: developers may be influenced by the conditions of their local or national political context in the form of policies introduced and pressures exerted by prominent political parties, lobbying groups, and government officials. Furthermore, prevailing political sentiments expressed by the public may also influence developers' actions, particularly recent concerns around the potential risks of AI technologies.
    - *Cultural factors*: while developers can be national or multi-national entities, they are embedded within particular cultures (e.g., ethnic, religious, national, corporate, industrial) with their own traditions and values that can influence production.
    - *Local factors*: individual programmers that operate independently to develop generative AI programs may be influenced by their immediate social relations such as friends, family, and local community. Companies and institutions may be similarly influenced by their official and unofficial links to other organizations including businesses, industrial networks, and trade unions.
- *Generative AI program:* the technical characteristics of the generative AI program itself, including the architecture and performance capabilities, will significantly influence the deepfake content produced. Notably, the technical conditions of the program may limit the accuracy of the content produced or restrict content to a particular format (e.g., audio, static image, audiovisual) or particular style of deepfake (e.g., human faces, geography).

Generative AI distribution

This grouping accounts for those social actors involved in the marketing, advertising and distribution of the generative AI program to the desired audience and/or general public. This includes the distributor themselves (e.g., marketing team) and also their outputs including marketing material, advertisements, and their use of online marketplaces.

- *Distributor*: once developed the generative AI program itself must be distributed to the developer's desired audience/market including the general public. The motivations and characteristics of those social actors involved in distribution may influence how the program and its

features/capabilities are communicated and in doing so influence how it is used by content creators. For example, marketing material for the program may encourage specific uses (e.g., entertainment, pornography) or be designed to appeal to particular user groups (e.g., social media influencers, programmers).

- *Marketing material*: how the program is presented in marketing material may influence who uses it for what purposes. For example, a program marketed as user-friendly will be more likely used by the general public and non-experts, while a program marketed on its technical specifications may discourage non-experts from using it.
- *Advertisements*: similarly to marketing material, advertisements featuring the program circulating online may encourage or discourage certain people from using the product, or further influence how they use it.
- *Online marketplaces*: how the product is presented and categorised within online marketplaces may further encourage or discourage certain people from using the product and influence how they use it.

Deepfake content creation

This grouping accounts for those social actors involved in the creation and publication of the specific deepfake content (e.g., text, audiovisual, audio). This includes the structure and composition of the training dataset used, the characteristics and motivations of the content creator (e.g., bad actor, entertainer), and the form of the deepfake content itself (e.g., technical quality, type of information), as well as any accompanying content (e.g., post text, disclaimers) added to the final publication.

- *Target training dataset:* generative AI programs are trained to generate specific deepfake content using a curated dataset pf images featuring the target and how these datasets are constructed can influence the content itself. Notably, the size of the dataset may determine the accuracy of the final product, while the variety of images (e.g., photographs of the target in different situations, or photographs of the target from a specific time period) may determine the application.
- *Deepfake content creator:* the deepfake content creator describes the actor that curates the target training dataset, uses the generative AI program to generate a deepfake of the target, and publishes it online for a specific purpose (e.g., disinformation). This may refer to an individual, a group of individuals, or an organization and accounts for various characteristics associated with the actor including, but not limited to, their demographic, political affiliation, organizational structure, and accessibility to resources.
- *Creator social environment*: the actions of deepfake content creators are also determined, to some extent, by various political, cultural and social factors at play in their local and general environment. As content creators may be understood as national institutions, multi-national corporations, or independent citizens, these factors may differ significantly.
    - *Political factors:* content creators may be influenced by the conditions of their local or national political context in the form of policies enabling or limiting access to generative AI technologies (e.g., bans on deepfake apps). Furthermore, as deepfake images present an opportunity for political disinformation, content creators may be influenced or incentivized by prevailing political sentiments and discourse to produce deepfakes that mimic political figures or that address political topics.
    - *Cultural factors:* content creators are imbedded within particular cultures with their own traditions and values that may influence the deepfake content produced, for example a patriarchal or misogynistic culture may normalize deepfake pornography targeting women.

  - *Local factors:* content creators that operate independently to produce deepfakes may be influenced by their immediate social relations such as friends, family and local community. Companies and institutions may be similarly influenced by their official and unofficial links to other organizations including businesses, industrial networks, and trade unions.
- *Accompanying content:* deepfake content published online will be accompanied by other content provided by the creator that frames the content itself and may influence how it is viewed. This includes titles, descriptions, web links to other content, advertisements, and special effects. Notably, deepfake content may be accompanied by explicit disclosures of inauthenticity.
- *Deepfake:* this describes the content of the deepfake content itself including the technical accuracy, the actions performed and words spoken, the overarching argument/sentiment expressed, and how the deepfake content is integrated with authentic content.

Target

This grouping accounts for those social actors related to the intended target of the deepfake content, primarily the characteristics of the target (e.g., demographic, public profile, political role) and those factors within their social environment that may contribute to their media representation (e.g., industry, political landscape, cultural values).

- *Target:* the target describes the actor represented within the deepfake content produced and may refer to an individual or group of individuals (e.g., public figures), an object (e.g., geographical landscape), a historical event, or a hypothetical scenario. This further accounts for the various characteristics associated with the actor including, but not limited to, their demographic, position and status in society, and their historical, political or informational significance.
- *Target social environment:* the various political, cultural and social factors at play in a target's local and general environment may make them more vulnerable to deepfake technologies. As targets may be understood as individuals, objects, or even events, these factors may differ significantly.
  - *Political factors*: a target's significance within the local or national political context may mean that they are more desirable targets. For politically significant figures (e.g., government ministers, officials) who make numerous public appearances/announcements, there may be more images of them available for use within a training dataset such that they are more easily mimicked and a deepfake of them will have a greater impact. For politically significant events (e.g., protests), however, the opposite may be true in that if fewer audiovisual recordings of these occurrences exist then a deepfake of the event may be more difficult to generate but also more difficult to debunk.
  - *Cultural factors*: targets are embedded within particular cultures with their own traditions and values that may render them more vulnerable and desirable to deepfake technologies. For example, figures seen as culturally significant (e.g., media celebrities) may be more desirable deepfake targets, while a misogynistic culture that sexualizes women may make them more vulnerable to deepfake technologies. Similarly, the cultural significance of landmarks, religious sites, and traditions/practices may mean they are more likely to be mimicked using deepfake technology as they may evoke stronger emotional reactions.
  - *Local factors*: a target's immediate social environment (e.g., friends, family, colleagues) may also make them more vulnerable or desirable as targets in that the

production of deepfakes may be more damaging to their personal life. For example, deepfake pornography may have further consequences for a target's family members and so it could be used to greater effect to intimidate the target and manipulating their actions.

Online circulation

This grouping accounts for those social actors operating within a specific social media platform and that mediate the dissemination of deepfake content from source to viewer. First and foremost, it is necessary to consider the architecture of the specific media platform (e.g., Twitter, Facebook) that dictates how users send or receive content via different information channels (e.g., newsfeeds, private messaging, advertising). However, other actors operating within this architecture play a significant role in the dissemination of content including content moderation programs, recommendation programs, platform policies and community standards, and other platform users sharing, reacting to and commenting upon, and remixing deepfake content.

- *Platform architecture:* different online media platforms enable users to interact with one another and disseminate content in different ways that may influence how deepfake content is received and consider by the viewer, notably how individual content is visually communicated to a user. As user interfaces differ between platforms the relevant characteristics may change but, broadly speaking, these include the compilation of personalized newsfeeds (e.g., TikTok's "for you" feed), the linking of related content via topic-specific metadata (e.g., Twitter hashtags), and the organization of content on the platform more broadly (e.g., grouping popular content in YouTube's "trending" section). Furthermore, how audience reactions and discussions on specific content are made immediately available to the viewer through visible metrics (e.g., recorded "likes", "dislikes", and "follows") and public forums (e.g., comments sections) may influence how content is received as public discussion may encourage either acceptance or scepticism. In relation to deepfakes, it is particularly important to analyse how inauthentic content is categorized on a platform and how it is presented or framed in relation to other authentic content such that its artificiality is obscured or de-emphasized.
- *Content recommendation programs*: within the established architecture of the platform itself, different AI recommendation programs dictate the flow of content between users by personalising what individual users see and do not see when browsing online. Depending on the program, recommendations presented to viewers may be based on their recorded preferences, their links to online communities and groups, their personal connections with other users, previous content they have viewed or interacted with and how they have interacted with it, and the content they have created, posted and shared themselves. As such, these programs often encourage homophily in that viewers are encouraged to only interact with users that are similar to them and content that agrees with their initial positions. In relation to deepfakes, recommendation programs may influence what kind of content is made available to a specific viewer and may also encourage viewers to interact with this content positively in that it is presented as similar to their own interests and in agreement with their values and the values of their communities.
- *Platform policies*: this actor describes the observable application of platform policies, standards and regulation that determine what forms of content are deemed acceptable on a specific media platform and therefore influence what content is made available to viewers and how it is made available. Such policies and their application may differ significantly depending on the platform (e.g., pornographic content allowed on certain sites but not others) but will likely include the masking (e.g., Twitter initially obscuring potentially sensitive content from view),

flagging (e.g., Twitter applying mis/disinformation labels to content), and the blocking of content. Notably, the application of these policies may be dependent upon national and multi-national regulation governing relating to, for example, mis/disinformation, sensitive content, and child protection and so is closely related to the viewer's social environment. In relation to deepfakes, it is necessary to analyse the impact of flagging or labelling deepfake content as inauthentic or AI-generated, particularly as this is a policy stipulated in the upcoming EU AI Act's transparency obligations.

- *Content moderation programs*: while the application of platform policies may be determined by moderation teams working for the platform company itself (these will be discussed further in the third party certifier section), more often automated moderation programs are used to filter and categorise content before being made available for circulation. As platform polices on content moderation are open to interpretation, technical characteristics such as the digital architecture and training datasets used will determine the actions of these programs and in turn will influence what and how content is made available to viewers.
- *Platform users*: deepfake content circulating online passes between numerous users that interact with it in different ways (e.g., liking, sharing, commenting) that influences how the content is distributed (e.g., trending content on Twitter will reach more users) and how the content is received by other users (e.g., positive reactions to content may encourage others to react positively also). These users may be motivated by political sentiments and beliefs such that even those that recognise deepfake content is fake may share and promote it online.
- *User community network*: other than recommendations and popularity, users will often receive content via members of their own immediate community network (e.g., friends, family, colleagues) who may promote deepfake content and aid in its deception or dismiss deepfake content as inauthentic and raise awareness of these issues. The actions of these members may significantly influence how users interact with deepfake content.

Public discourse

Deepfake content does not exist online in isolation but rather relates to other pieces of content and the relationship between the two may influence how the deepfake is received and spreads online. Notably, the production of deepfake content is informed by and subsequently contributes to various prevailing media narratives that the viewer receives from content published in the broader media landscape outside of the social media environment (e.g., news stories, artistic representations). This segment describes those prominent media narratives that might influence the viewer including those relating to the content creator (e.g., trustworthiness, political affiliations, expertise), the target (e.g., public persona), the generative AI developer (e.g., trustworthiness), generative AI technology more generally (e.g., capabilities, safety), and potentially the topic addressed within the deepfake content (e.g., controversy, disputed).

- *Media narrative of source:* news outlets and online publishers similarly have established media identities such that their content may be viewed as more or less trustworthy, politically biased, or otherwise considered differently (e.g., comedic or satirical sources). The media identity of the source may encourage viewers to more readily accept or question the authenticity of deepfake content.
- *Media narrative of target:* deepfake content often features prominent public figures associated with a particular media persona that has been artificially established through a wide composition of online content (e.g., articles, public statements, media appearances, interviews). However, every social media user constructs a personal media narrative through their public posts, photographs, and relationships online. If the deepfake appears incompatible with the

target's media persona, viewers may be more likely to question its authenticity or it may seem to cause more significant reputational damage.

- *Media narrative of AI:* one specific media narrative to consider in detail is that related to AI technology and deepfake content itself. For example, one media narrative may downplay the accuracy of deepfake content and therefore encourage uncritical acceptance, while another narrative may overemphasise the accuracy of deepfake content and therefore encourage extreme scepticism. This is of particular concern when considering the public opinion and uptake of potentially positive and pro-democratic uses of AI technologies.
- *Media narrative of topic:* as the targets are often high-profile public figures whose statements and opinions carry significant influence, deepfake content is linked to particular social, political or cultural issues that are presented to users through different media narratives. These media narratives may influence how people receive deepfake content but simultaneously this content also contributes to the narrative itself.
- *Media narrative of developer*: the brand or identity of a generative AI developer can be established through company advertising and communications, public appearances and interviews with prominent members of the organization (e.g., OpenAI CEO Sam Altman appearing at congressional hearings), and stories circulated within broader media. Such communications may contribute to perceptions of a developer's intentions, competency and trustworthiness which could influence how people view the use of their technology or content generated using such technology including deepfakes.
- *Media narrative of the platform*: social media platforms themselves have brand identities that may influence how users may view the content circulating on them. For example, content shared on platforms with strict policies on mis/disinformation and effective moderation programs may be viewed as more likely to be authentic, while content shared on platforms with no moderation may be viewed as more likely to be inauthentic. The platform's identity in terms of safety may be influenced by advertisements/promotions issued by the platform company itself or further through media reports on platform policies or high-profile cases.
- *Media narrative of technology:* advertisements and marketing material promoting the generative AI program itself, as well as media reports, may establish a perceived identity of the product that may influence how users view its use. For example, media reports emphasizing inaccuracies and bias in ChatGPT's responses may cause users to view the program and its outputs as unreliable.

Policy and legislative intervention

This grouping accounts for those interventions from policymakers at different stages of the process of production, circulation and reception. The actors involved include policymakers themselves (e.g., government officials, elected politicians, regulators), specific AI legislation (e.g., EU AI Act), trade and marketing regulations (e.g., transparency obligations), platform regulation (e.g., GDPR) and certification (e.g., fact-checking, pre-bunking).

- *Policymakers:* this generic term refers, in this case, to those involved in the creation and implementation of policy, legislation and regulation that directly intervenes in the spread of deepfake content. The characteristics of these policymakers (e.g., political affiliation, social status) may influence the form of these policies and how they are implemented.
- *AI legislation:* though there is legislation addressing harmful online content, AI programs pose specific risks that require specific legislation such as the EU's proposed AI Act which specifies transparency obligations for generative AI that requires AI-generated content to be labelled as such. Further AI legislation may come to be introduced in the next few years and

such legislation may further influence how generative AI programs are developed if particular uses of this technology (e.g., production of deepfakes) are deemed to be of risk.

- *Trade and marketing regulation:* regulations and laws imposed upon the development of AI technologies may differ from those imposed upon the marketing of such technologies and thus influence how generative AI programs are distributed to the wider public. For example, regulations preventing misleading advertising and marketing material can prevent distributors from misrepresenting AI programs and their capabilities.
- *Platform regulation:* regulations governing the spread of information on social media platforms will also intervene in the distribution and circulation of deepfake content. For example, certain regulation may prevent pornographic content from being circulated on certain social media platforms and/or to specific users (e.g., children) and thus influence how deepfake pornography spreads and to whom.
- *Certification:* numerous independent news groups and organizations operate as fact-checkers on social media platforms identifying and correcting false or misleading claims and content circulated online. These interventions could be used to identify deepfake content and work against media narratives that said content is communicating such that they may influence how such content is received. Fact-checkers may label particular items of content as false or misleading, as well as sensitive, offensive or unverified. While these interventions may not block or alter the visibility of the content itself, they may influence how users consider such labelled content. Furthermore, pre-bunking initiatives are used to raise awareness of media and discourse manipulation techniques in order to improve critical thinking and media literacy. The spread of pre-bunking material may enable viewers to identify deepfake images and critically engage with the narratives they progress.

User reception

This grouping accounts for those social actors related to the viewer of the deepfake content, primarily the characteristics of the user (e.g., demographic, media literacy, political affiliation) and those factors within their social environment that may shape their perceptions or understanding of the deepfake content (e.g., political landscape, cultural values, social group).

- *User*: an individual social media user viewing deepfake content may tend to trust or distrust this content based on a number of personal characteristics including, but not limited to, their demographic, education level, media literacy, knowledge of AI technologies, and political affiliation. Furthermore, these individuals may hold specific roles within society that could leave them more vulnerable to susceptible to targeted manipulation through deepfake technology as their actions may have a wider impact such as journalists, government officials, academics, and military personnel. Individuals
- *User social environment*: various political, cultural and social factors at play in a viewer's local and general environment may make them more susceptible to deepfake content.
    - *Political factors*: viewers may be influenced by the conditions of their local or national political context in the form of policies and regulation, local pressure groups, and prevailing political sentiments expressed by the general public.
    - *Cultural factors*: viewers are imbedded within particular communities and cultures (e.g., ethnic, religious, national), as well as those that exist within the institutional structures in which they work (e.g., civil service, military, academia). The traditions and values of such cultures may influence how viewers receive deepfake content in that they may be more susceptible to or sceptical of particular content. For example, journalists may be more willing to accept and promote apparently scandalous content featuring politicians in order to appeal to a wider viewership.

- *Local factors*: individual viewers may be influenced by their immediate social relations such as friends, family, and local community, as well as colleagues and superiors within their organizations. Furthermore, their access to detection technologies may influence their reception of deepfake content.

2.2.1 Actors and Relations

The theoretical framework of ANT and of socio-technical systems help us lay down which actors are present, as discussed above and shown in Figures 2.1 and 2.2., but the nature of the *relations* between humans and machines/instruments/artefacts/technologies are continually changing and so difficult to describe in detail. This is a core question of philosophy of technology and of philosophy of information and one that has been tackled in two influential approaches. In this section, we first present these approaches and then explain why, although building on them, the approach we take in SOLARIS is significantly different.

An important approach to consider is *post-phenomenology*, that identifies the several different types of relations between humans, artifacts and the world around us. Ihde's (1990) original four types are as follows:

- Embodiment relation: (human—technology) → world
  - Human experience is reshaped through technical artefacts (e.g. wearing eye glasses)
- Hermeneutic relation: human → (technology—world)
  - Humans learn / adapt to give meaning through interactions with artefacts (e.g. using a wristwatch, or nowadays a smartwatch)
- Alterity relation: human → technology(—world)
  - Humans relate to technologies as quasi-others (e.g. the way in which we interact with Alexa or Siri)
- Background relation: human (—technology—world)
  - Technologies become part of the 'normal' background of humans (e.g. nowadays it is normal to have a fridge, at least in the global North and Global West)

Verbeek has made further additions to Idhe's original types that are specifically about intentionality: (Verbeek, 2011):

- Cyborg relation: (human/technology) → world
  - The intentionality of humans blends with that of technologies, e.g. using devices for hearing impairment
- Composite relation: human → (technology → world)
  - The intentionality of the human adds to the ones of technologies, e.g. certain instruments have a different way of relating to the world than us.

Note that the arrows indicate the (kind of) intentionality or directedness at play between the human and technological entities in each case, while the dashes indicate "a relation between entities which is not specified further" (Verbeek, 2011, p. 143) and can be thought of descriptively or theoretically as "black boxes" of the intentionality between the entities (ibid.). Furthermore, the parentheses indicate that either the human, artifact or the world is isolated from the rest of the schema.

The second approach to explaining human-technology relations derives from the philosophy of information in which Floridi (2014) has introduced the concept of 'in-betweenness' and the 3 types of relations.

'In-betweenness' refers to the different forms of interactions between agents, qua users, and technologies. The general scheme involves a 'prompter', i.e. whatever stimulates or suggests using a given technology, a user, and a technology in between them:

user ← technology → prompter

The simplest type of relation is called *first-order technology*, in which technology is interposed between humans and nature:

humans ←technology → nature

First-order technologies are, for instance, sunglasses to protect our eyes, an ax to split woods, or spectacles to improve our vision. In science, we can think of a simple meter or an analogue telescope to be this kind of technology.

With *second-order technologies*, we alter this configuration, and now technologies are in-between humans and another technology:

humans ←technology → technology

Thus, for instance, we use a screwdriver to operate on a screw, or a remote controller to turn on the TV. Many household appliances or scientific instruments before the digital revolution can be of that type.

Finally, *third-order technologies* are those in which humans are (allegedly) out of this chain, according to the scheme:

technology ←technology → technology

Numerous digital technologies can interact with other technologies without humans being present. Prima facie, this happens in the internet of things, in deep-learning algorithms, or in 'nested' climate models.

With either approach, we can consider the type of relations between humans on the one hand and artefacts on the other hand, but in isolation. As a consequence, if we aim to understand how human users come to 'trust' AI-generated contents, we will look at specific, individual characteristics of the artefact, for instance quality of video or of sound. SOLARIS, however, conjectures that how and why human users come to trust AI-generated content needs a more subtle analysis. In this deliverable we explain the theoretical foundations of this conjecture, which will be subject to empirical study in WP5.

The new approach advanced by SOLARIS integrates the concept of the "poietic agent" introduced by Russo (2022) with the ANT analysis presented above. Russo's poietic agent extends Floridi's concept of "homo poieticus" so as to properly describe the poietic powers that human and artificial agents have. Of particular interest to the project are the poietic characteristics that she ascribes to artificial agents. Russo's argument follows two primary trajectories.

Firstly, Russo argues that artificial agents qualify as genuine agents due to their ability to process information, a point already made in the Philosophy of Information (Floridi, 2014). This capability transforms artificial into *epistemic* agents, which can alter the dynamics between other agents and their environment. Russo's perspective on epistemic agency is inherently relational. Knowledge, in her view, emerges (inter alia) from the collaboration of all agents capable of processing information. These agents can influence the environment, introducing new dynamics between it and other agents.

Secondly, to elucidate the partial autonomy and consequent "agentiality" of technical objects, Russo engages with the works of both Latour and Gilbert Simondon. Russo's stance on epistemic agency is anchored in constructionism, a philosophy that bridges the gap between realism (where knowledge mirrors an objective external reality) and constructivism (where reality is crafted by epistemic agents) (Floridi, 2014) Specifically, a consequence of constructionism for knowledge is both relational and distributed among the epistemic agents within a socio-technical system. This implies that knowledge is always contingent upon a specific configuration of relations—both between the world and the epistemic agents and among the agents themselves. Therefore, evaluating the epistemic validity of a concept or claim is contingent upon the "level of abstraction" at which the analysis is conducted. Russo explains the idea of 'relational knowledge' in the following terms:

"Knowledge is relational also at the level of the concepts or of the semantic artefacts that compose it. I speak of semantic artefacts to emphasise that we make the concepts that we use to make sense of the world around us. I explored the idea that human epistemic agents are makers not just because we make artefacts, but also because we make semantic artefacts or concepts. To say that knowledge is relational at the levels of concepts means that these are not islands but are always connected to other concepts. I take this to be an irreducible relational aspect of knowledge." (Russo, 2022)

The pivotal insight of this conceptual framework is the notion that knowledge, including concepts, is a kind of artifact that stems from the epistemic collaboration between poietic agents, both human and non-human. It is this core concept that SOLARIS advances in its approach to analysing the production, circulation and reception of AI-generated content as the poietic agent allows us to rethink our understanding of artificial agents and enables us to address the question how human users come to trust AI-generated content . As explained in detail earlier, trust in AI-generated content is not a function if their technical features only. We need instead understand the broader network of relations in which such content is developed and generated, and further disseminated and spread. The network is socio-technical, but we also need to recognize that the 'artificial' actors present in this network are not just instruments through which AI-generated content is shared, but are proper agents that human agents interact *with*.

The consequences of this theoretical approach will have important impact for formulating policy options: following the current framework, policies will focus on regulating technical aspects of deepfakes, or in any case issues pertaining to how the technological artifact is made (for example, obligation of transparency), and will not focus on the *interaction* between the human and non-human actors, that will shape the social system in any case. If we want to regulate well this field, we need to understand the relations between humans and technologies from *a system* perspective, as this may give us various and numerous leverage points. In this deliverable, we further develop the empirical set up to test our theoretical framework. In WP5, we run three use cases based on these theoretical premises, and in WP4 we develop the consequences for the regulatory framework.

2.3. A Visual Semiotic Approach

In this section, a semiotic and visual analysis of the deep fake is explored. This is complementary to the ANT network presented above. Below we draw the connections between the network and a semiotic approach. A general assumption is that deep fake image or video is never alone. As showed above, the enjoyment of deepfakes is facilitated by social discourses that surround them. Deepfakes are not isolated, rather they are a part of communicative flows. They have an internal context in which they exist and a broader external context in which they may have been embedded. They can also be linked to other textual objects in a cause-and-effect relationship.

Images in deepfakes are organized on multiple semiotic levels. For instance, deepfakes combine visual and verbal language, making them syncretic. Deepfake videos, on the other hand, combine visual, verbal and sound languages.

Circulation/User Reception: The first interpretative step is establishing which social discourse these texts belong to according to the *social field* in which they spread.

Social discourse means the text's set of topics and objectives (e.g. satire, denunciation, enquiry, scientific research, revenge porn etc.) (Greimas, 1976).

The *social field* is the pragmatic scenario in which the text is released (e.g. a museum, health services, a criminal trial, social media, a website, an App, television, cinema etc.). Sometimes, the field can guide the interpretation of the relevant discourse, but there can also be collisions between the field and the discourse.

On 24 March 2023, a Reddit user posted a fake picture of the pope wearing a Balenciaga puffer jacket via Midjourney. The photo's discourse is satirical, but the social field in which it is released is that of a mixed, partly journalistic platform that offers news content. The photo went viral and was interpreted by many Reddit users as authentic. An analysis of the reception shows that there have been three main interpretations of the image: (1) as a meme; (2) as a sinful action, regardless of the truth value of the photograph; (3) as a real photo, explanatory of the Vatican's excessive wealth. Reception analysis (Geninasca and Sadoulet, 2000) assumes these clashes as analytical chances to clarify which cultural competencies are needed for different deep fakes. Reception analysis takes that various deepfakes are recognisable concerning specific social discourses/fields.

Public Discourse: Other known deep fake cases have generated controversy (Raynaud, 2003; Pinch and Leuenberger, 2006; Reber, 2006). A solid ethical dimension concerns deep fakes of deceased people. In 2021, the Spanish beer company Cruzcampo issued a deepfake of the Andalusian singer Lola Flores, who was used as the star of an ad video. Flores died in 1995. Although Flores' family supported the campaign, this triggered an outrageous reaction. The video went viral, while the Spanish debate focused critically on the abuse of the memory of a cultural icon for commercial purposes. The example highlights several elements. Firstly, the viewers of the video were aware that they were watching a clone of the star. Secondly, the impossible direct consent of the singer was not at the centre of the debate, unlike in other cases.

Thirdly, the technological innovation that enabled the deep fake was not perceived as a tribute to the singer, as Cruzcampo had claimed, but as an unfair exploitation for economic reasons. Thus, the semiotic value[1] of *tradition* and that of *innovation* are contrasted in the controversy. Tradition is interpreted positively as a collective value in which the whole community participates, while innovation is seen as an individual and undue process, not helpful in preserving the collective cultural memory (Marrone, 2014, 2021).

The matter can be compared with another well-known debate. In 2021, a controversy arose over a deep fake of a human voice. Filmmaker Morgan Neville directs a documentary about US chef and writer Antony Bourdain, who died in 2018. Needing a clip of Bourdain's voice to use in the film and because Bourdain was no longer alive, Naville arranged for the deepfake to be made. Naville

[1] The term *value* has itself several meanings. We differentiate, for example, between value understood through 'valuation', or estimated worth or price, and value understood as 'quality' which makes someone or something worthy of esteem, desirable or important. Semiotic theory describes value as arising from the relationship between actantial subjects and objects: any subject's need or desire for a particular object makes the latter valuable, turning it into *an objet de valeur* in the process. Moreover, the value it has for the subject comes to be identified with the object. For instance, if someone buys a car, it is probably not so much a question of owning the object/car but rather of acquiring an easy and comfortable means of transport, or a way of enhancing one's social reputation, or enjoying a feeling of power. The thing itself, in this case, is merely pretext, a placement for the desired values. Thus, in semiotic analysis, the term object of value has been fashioned to designate objects placed in relation to subjects. Social processes and discourses can be thus conceived as for a flow of circulating semiotic values (Martin, Ringham 2000).

consulted Bourdain's widow and his literary agent, and they did not object. However, there are two different elements from the Cruzcampo case. The documentary's viewers were unaware that they were listening to a clone. Moreover, one of Bourdain's relatives protested against this choice of director. The story became a major news story, triggering a debate on documentary film as a film genre sensitively devoted to truth, media manipulation, and the 'right to one's voice'.

Lola Flores and Anthony Bourdain are both public figures and artists. Still, in Bourdain's case, the controversy is not based on the contrast between *tradition* vs. *innovation* but rather on the values of *privacy* and *respect for the individual* as opposed to *impersonal* and *untrue technology*.

The two controversies are based on the perception of generative A.I. as a kind of 'bad joke' (at the narrative level) with unpleasant effects. Still, in the Spanish case, the outrage seems specifically related to cultural anger, whereas in the US case, it is a voyeuristic gesture. When referring to values or narrative elements, semiotics conceives them about a general model:

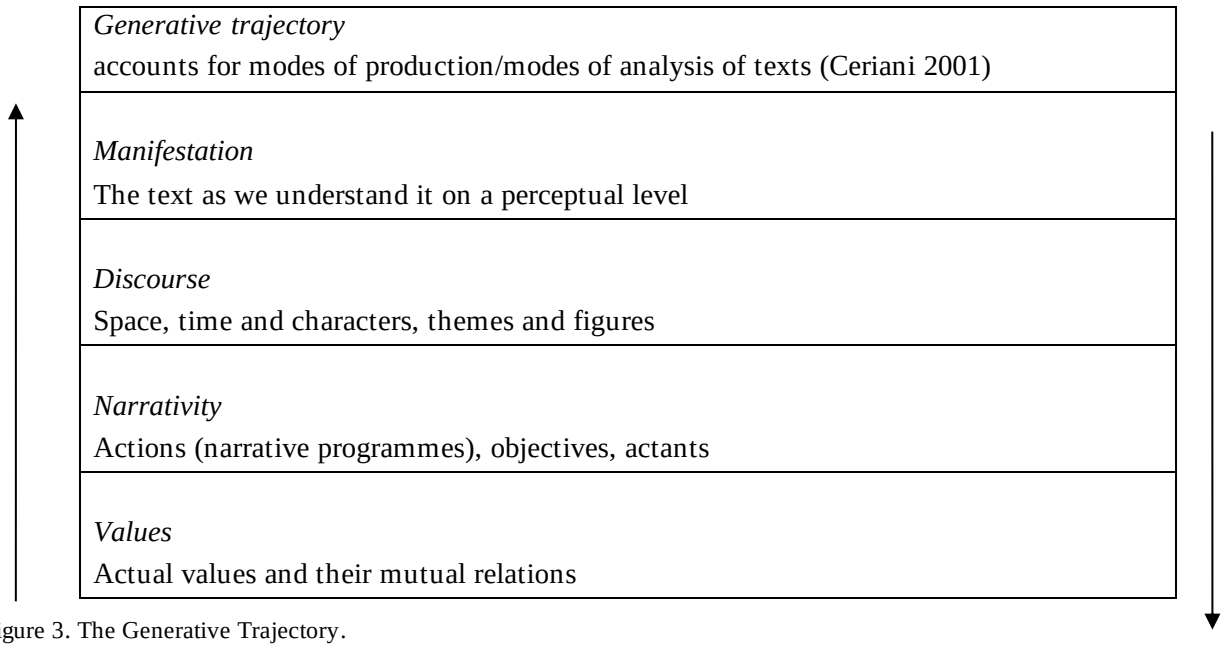

Figure 3. The Generative Trajectory.

The controversies allow us to draw a picture of the ethical values, translatable into semiotic values, at stake in the case of deep fakes. These analyses are preliminary to designing the effectiveness of the possible positive uses of deep fakes.

Target: Eco (Eco, 1976, 1991) made a relevant contribution concerning communicative flow and users' interpretative processes. Eco's approach must be presented separately because it constitutes a specific approach. Eco's approach helps understand user behavior on social media, for example, and accounts for the 'first' reactions we cognitively bring into play to interpret something as a deepfake video via TikTok, Telegram or Twitter.

Eco's approach is based on two assumptions.

(a) Users always collaborate in constructing the meaning of an image, video, sentence or film, i.e., a text in general. This user/viewer/reader fills in the many gaps in the text, which is only partially explicit, using anything from simple linguistic inference to more complex deductive reasoning that applies to the entire meaning-making process.

(b) Just as there are different discourses and social fields, there are different communities regarding competence. Interpretive competence, in other words, does not only change from one individual to

another but must be conceived as a capacity of social subgroups. This is increasingly important for synthetic media languages, which evolve very rapidly. In the face of these languages, the distinction between young and mature generations, or between Millennials, Generation Z and Generation Alpha, needs to be stronger and generalising.

User Community Network: To define how we interpret something new, such as a reel we find on our smartphone in a continuous video stream, Echo proposes to distinguish three aspects, which can be analysed separately and in detail.

- *Cotext.* What might be before or after the reel we are watching? Cognitive expectations are formed concerning the virtual organisation of a communication flow. Deep fakes via social media are often perceived as memes.

- *Context* (or *frames*). The socio-cultural information we own to 'predict' and accept the actions in a video. The report lets us understand that we are watching an entertaining video, not a political rally, or conversely, the latter and not the former.

It seems possible to approach context with the tools of visual semiotics (below).

- *Circumstances.* The social field and discourse were mentioned, but Eco's notion of circumstances is more specific. Alternatively, instead, it can help articulate many types of technological interactions that the concept of the social field cannot define. Circumstances answer several questions: in which 'situation' do we approach the text? On what medium? With what accessibility? Who sends it to us, and why (Bassano, Chessa and Solari, 2023)? For example, deep-fake pishing or catfishing through deep-fakes occurs under particular circumstances. These phenomena are not only defined by the strategies of the swindlers and the competencies of the people frauded but also by a specific interaction process that the Eco's notion of circumstances allows us to focus on.

Target and Cultural Factors: a visual semiotic approach

Visual semiotics, as explained by Polidoro (2008), is a field of semiotics that focuses on visual and audiovisual texts and perception. It is typically divided into two areas: figurative semiotics, which deals with meanings and interpretations derived from recognizing objects in an image, and plastic semiotics, which is interested in how visual configurations are inherently meaningful. This approach helps us understand the various levels of meaning involved in evaluating a visual or audiovisual text and determining its authenticity. In other words, interpreting an image depends not only on its perceptual features but also on cultural mechanisms. When examining visual fake news and deep fakes, the primary concern of Visual Semiotics is to understand how the content of an image can influence the perception of its trustworthiness.

Enunciation and Utterance

There is a distinction between *enunciation* and *utterance* (Greimas and Courtés, 1986). Utterance is the level of the message as a communicative product. *Enunciation* is the set of traces, more or less explicit, of the fact that the text has been produced. The trustworthiness of enunciation is autonomous from the trustworthiness of utterance. Sometimes, the truthfulness of a text, image, or video depends on these two levels' collaboration. However, often, there are clashes between enunciation and utterance.

Let's discuss one of the most well-known instances of GANs-generated deep fakes - the TikTok videos from the @deeptomcruise series in 2021-2022. Fast forward to 2023, the series has gone viral, and differentiating between these parodies and an actual Tom Cruise clip has become increasingly challenging. In one of the parodies, the actor is seen dancing to a remixed Lady Gaga song in a typical TikTok challenge, while wearing a bathrobe in a garden outside a private home. We can evaluate the

truthfulness of the message (utterance) and the way it is presented (enunciation) separately. Let's consider Tom Cruise participating in a TikTok challenge and sharing it with his followers. Is it likely to happen? Is the background of the video consistent with a genuine setting? Does the person in the video perform the dance that has become famous through TikTok challenges in a believable way? Could that person be Tom Cruise? All these aspects concern the message itself or what the video is conveying. The answer to these questions is mostly positive: the message is coherent, the person in the video looks like Tom Cruise, and it is plausible to think that Tom Cruise might participate in a TikTok challenge as a user of the platform.

It is important to note that there is an issue with the video's enunciation. Tom Cruise's face looks young, as if he were around thirty years old. This is a problem because Tom Cruise is currently 59 years old in 2021, while TikTok was founded only seven years ago in 2014. The reason for this discrepancy is that the GAN network created the deep fake by analyzing many hours of films from the 1990s and 2000s, when Tom Cruise was in his thirties. Therefore, the incompatibility between the age of Tom Cruise's face and the founding date of TikTok provides evidence that the utterance may not be genuine.

Verbal anchoring

At the level of *enunciation*, a second important aspect of the Visual Semiotics approach is the focus on the relationships between verbal, visual, and possibly sound parts of deep fake texts. As mentioned above, a text such as a deep fake is always organised by the simultaneous presence of verbal and visual features.

In semiotics, many studies have been conducted on how the verbal text orients the visual text, 'anchors' it, and guides interpretation (Floch, 2000). This applies to the subtitles in a video, to the logo of the channel or social platform disseminating the video or image, and it is crucial to carefully design the verbal parts of deep fakes for their effective use towards AI for good.

Note: The 'Balenciaga Pope case discussed above shows a remarkable short circuit. As mentioned, the video was released on Reddit. After a few hours, it was posted on Twitter, but the viral spread had already begun. After the image of the Pope had already gone viral, a note was appended to the original tweet that shared it: “This image of Pope Francis is an AI-generated picture and not real”, the note reads. “The image was created on the AI image-generating app Midjourney”. In this case, since the visual text had begun to circulate without this verbal part, some irreparable damage was done so that the note was deemed by many to be irrelevant to interpreting the image.

In the *Malaria Must Die* campaign, gestural and sound credibility effects are kept separate. David Beckham's lip movement is modified, but not the original timbre of the native speaker actors' voices. This does not weaken the effectiveness of the text but instead makes it more transparent as AI-generated content.

Figurative and Plastic Isotopies

Returning to the utterance level, a deep fake's trustworthiness must be assessed along two complementary paths. On the one hand, there must be a consistency of the *plastic* elements: the high technical quality that makes a face realistic, the lighting, and the details of the clothing in the case of the pope's deep fake, must concern, as noted by many, all the plastic components of the image. Thus, the vaguely outlined crucifix and the pope’s right hand, a strange curve in the outline of Pope Francis' glasses, make an experienced eye able to detect the fake. However, there is also a crucial *figurative* component, namely the coherence of the elements of the world that are included in the image. This coherence, which in semiotics is considered redundancy and is called *isotopy*, is determined by

cultural patterns. The background elements of the image or video must be 'plausible' in every detail, and this often makes it relatively easy to examine a deep fake by detecting oddities and inconsistencies.

3 Setting the psychometric scale

3.1. Background and aims

One of the project's aims is to develop a psychometric scale capable of measuring the perceived trustworthiness of GANs-generated content that captures whether users perceive such contents as trustworthy (i.e., as truthful and not fabricated). At the moment, there are no valid, reliable, and comprehensive scales of perceived trustworthiness of GANs-generated content which take into account a variety of aspects (e.g., those related to the person on the video, source, message, and other aspects) that are important when making conclusions regarding deepfakes and other similar kinds of content. The lack of such measures is hindering the development of a more nuanced understanding of what drives such evaluations.

As such, we aim to design a valid and reliable multidimensional scale capable of assessing different features that enhance or reduce the perceived trustworthiness of GANs. In other words, the scale is focused on individuals' perception of content-related aspects (as opposed to individual characteristics, which will be captured with existing measures), which determine the likelihood that GANs-generated content will be perceived as trustworthy.

We want to include at least three perceived features of GANs, such as features related to the person on the video (e.g., vividness), source (e.g., perceived source credibility), and the message (e.g., information believability), although the specific number of dimensions was not set a priori; instead, the final number of dimensions was established during scale development (e.g., after item generation interviews and a thorough review of existing measures).

Our initial aim has been to develop the scale in at least two languages, but we have later extended this with an additional language. Hence, the current version of the scale is available in English, Slovene, and Italian. It will be used in further studies within the project, particularly in use case 1. Still, it will be made publicly available (through publications and conference presentations) for use in studies outside of the project as well.

Scale development procedure

The scale development procedure is aligned with the proposal but it is slightly expanded due to recent developments in psychometrics. First, we have conducted a literature review to identify relevant aspects that constitute the perceived trustworthiness as well as the questionnaires and items from existing studies that measure them. Second, we have generated an additional item pool via interviews with students and online surveys with other stakeholders, in which they have been asked to think aloud while attempting to decide whether GANs-generated content could be trusted. Relevant thoughts have been transformed into items and analyzed using thematic analysis to identify aspects commonly considered by the individuals. Third, additional items have been generated using Psychometric Item Generation (PIG) (Götz *et al.*, 2023), a natural language processing algorithm based on the GPT-2 that produces human-like customized text output. All sources of information have been then combined to generate a larger pool of potential items, and we made decisions regarding the proposed dimensions of the questionnaire.

Once we prepare the item pool, we ask the experts to assess the content validity of the collected and generated items, with the aim of shortening the scale, identifying the most relevant items, and adapting the items to make them as clear as possible. The resulting version of the scale is then translated using the translation-back translation technique to Italian and Slovene. This version will

later be used in a cross-sectional survey conducted in three countries, followed by psychometric analyses related to factorial structure, validity, measurement invariance, and internal reliability, and adapted according to our findings to achieve appropriate psychometric properties. All steps are outlined in Figure 4.

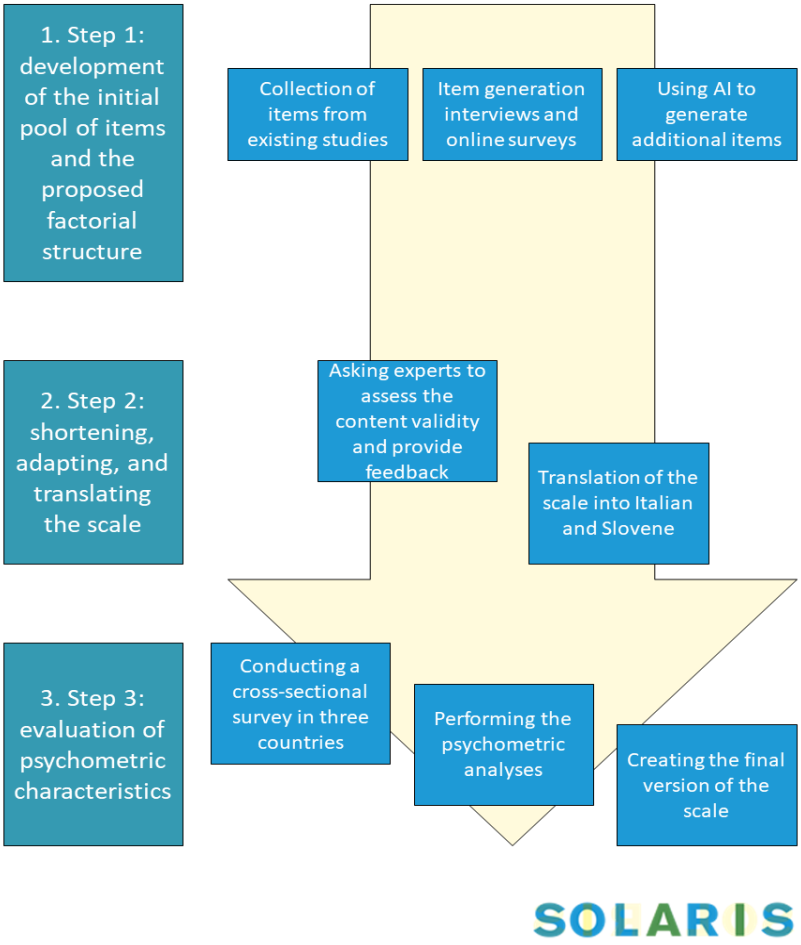

Figure 4. Scale development procedure.

3.2. Summary of the scale development procedure and findings

Step 1: Development of the initial pool of items and the proposed factorial structure

We have used various sources to develop the initial pool of items. First, we have reviewed the existing literature to collect items from existing scales and generate items based on aspects identified as relevant in previous studies. Second, we have conducted face-to-face interviews and an online survey, in which participants have been asked to think out loud while deciding whether they would trust the videos they were exposed to. Third, we have used the Psychometric Item Generator – a generative language model, based on the GPT-2 (Götz *et al.*, 2023) to further complement our pool of items. Each of these phases, as well as the synthesis of different sources, is described below.

Collection of items from existing studies

In the first phase, we have ran multiple searches in academic research databases, such as SCOPUS, Web of Science, and Google Scholar, to identify articles that explore individuals' perception and, in particular, trust of deepfakes and other manipulated videos. While most articles employ simple questions with a dichotomous response format (e.g., "Would you trust this video?"), we have managed to identify six research papers (Hwang, Ryu and Jeong, 2021; Hameleers, Van Der Meer and Dobber, 2022, 2023; Lee and Shin, 2022; Shin and Lee, 2022; Ng, 2023) that use scales with multiple items (i.e., two to five) to measure different aspects of deepfakes. Previous scales mostly focused on content (e.g., credibility of presented information, persuasiveness of presented information) and, to lower extent, the person on the video (e.g., vividness). Other aspects, such as technical features, were mostly neglected. In fact, the only existing scale that incorporates different aspects, is the one by Hameleers and colleagues (2023) (believability of the message scale). It asks participants about the content of the political speech, the political actor, and the way the message is presented (one item per aspect). This basic structure (content, person in the video, other aspects) has been used as a starting point for our scale.

Our literature review additionally yields three articles (Tahir *et al.*, 2021; Thaw *et al.*, 2021; Hameleers, Van Der Meer and Dobber, 2023) that do not use any relevant scales, but provide other valuable information. Specifically, Hameleers and colleagues (2023) analyze open-ended questions about trust, distrust, credibility, and believability (as they pertain to the message, source, and presentation in the context of politics) and identify five crucial topics: 1) a content-wise discrepancy between the political reality and the statements voiced in the message, 2) perceived manipulation and doctoring of the audio-visual stimuli, 3) broader levels of political distrust and opinion-based disagreement, 4) lack of factual evidence, sourcing, and argumentation, and 5) self-perceived media literacy. Topics 1, 2, and 4 are important in the context of our scale, whereas topics 3 and 5 refer to individual characteristics. Tahir and colleagues (2021) use eye-tracking and surveys to identify features that help identify deepfakes. They provide a list of such features, including, for example, blurriness on the eye region, subtle color differences, and clear boundaries of the face. Lastly, Thaw and colleagues (2021) conduct interviews and extracted the features associated with accurate deepfake identification. Similarly to Tahir and colleagues (2021), they provide a list of such features, for example lack of emotions, abnormal mouth movements, and unnatural voice. These aspects, identified in previous studies, have been used to generate items, with resulting items mostly referring to person-related aspects, but, to a lower extent, also to content-related aspects and other features.

Although literature on the perceived trustworthiness of deepfakes is just emerging (with all relevant studies being published in recent years) and the existing scales are relatively rudimental, previous literature does offer some insight into the process that takes place within the individual when deciding whether to trust GANS'-generated content. Altogether, our literature review has led to 41 unique items adapted from previous scales or generated based on previous studies, with 19 of them referring to the content, 18 to the person in the video, and 4 to other aspects (mostly technical ones).

Item generation interviews and an online survey

In the second phase, we have conducted face-to-face interviews with students (in Slovenia and Italy) and an online survey (internationally) aimed at three groups of crucial stakeholders, specifically citizens, journalists, and experts. The interviews and the online survey have had a similar procedure. In particular, individuals have been first given basic information about the study and asked to sign the informed consent form. They have been then exposed to four videos (three deepfakes and one authentic video depicting a famous person), shown one-by-one. Participants have been encouraged to share all the thoughts that popped into their head while watching the video, regardless of their relevance, and told that they will be asked whether they would trust each of the presented videos. At

first, individuals were not asked any specific questions, but, in the later parts of the interview/online survey, they were shown the videos once again and were asked what they thought about the content, person in the video, and other aspects. In the online survey, they provided their answers via text boxes (in English), whereas interviews were recorded, transcribed, and translated. At the end of the study, participants were asked to fill out a short demographic questionnaire. They were also debriefed that some of the presented videos were manipulated. Both interviews and the online survey took approximately 25 minutes to complete. During the analyses, participants' thoughts were organized into meaningful units (specific statements). All statements were then coded using deductive thematic analyses, in which each of the meaningful units was categorized into one of the three themes that emerged in the first phase (content, person in the video, other).

16 students from Slovenia (50.0%) and Italy (50.0%), all proficient in English, participated in face-to-face interviews. Most of them (75.0%) were women and the average age was 23.69 years (*SD* = 3.93). Half of them (50.0%) had previously completed a higher secondary education, 25.0% had previously completed a bachelor's degree, and the remaining 25.0% had previously completed a master's degree (and were PhD students at the time of the study). The majority of participants (56.3%) were psychology students, while the rest studied sociology, management, finance, economics, philosophy, and foreign languages. The interviews generated 82 statements pertaining to the content (e.g., "I would not say it is the truth. It seems like he is trying to impress someone with this story, so he is probably making this up."), 121 statements pertaining to the person in the video (e.g., "His gestures were off, it seems very unnatural to explain something in such a theatrical way."), and 303 statements pertaining to other aspects (e.g., "This was posted by a random person.").

10 participants (50.0% male, 50.0% female), all proficient in English, filled out the online survey. They were highly educated (80.0% with at least a master's degree) and their average age was 35.6 years (*SD* = 9.31). Half of the sample (50.0%) consisted of experts, while the remaining participants were either journalists (30.0%) or regular citizens (20.0%). Altogether, the online survey led to 40 statements pertaining to the content (e.g., "The content is unconvincing and contrived."), 70 statements pertaining to the person in the video (e.g., "The lips were not completely in sync with the text."), and 87 statements pertaining to other aspects (e.g., "Video seemed poor quality, low resolution.").

In the last step, we have then developed the items from unique and relevant statements collected via interviews and the online survey. The two think-aloud approaches together led to 305 items, with 43 referring to content, 87 referring to the person in the video, and 175 referring to other aspects.

Using a generative language model to generate additional items

In the third phase, we have used the Psychometric Item Generator. This novel open-source generative language model runs on Google Collaboratory and has previously been found to be an effective machine-learning solution to developing items for psychometric scales (Götz *et al.*, 2023). Specifically, we have followed the procedure outlined by its' developers; we have used the pre-trained 774M GPT-2 language model, have fed the algorithm 10 previously-developed items per each category (content, person, other), set the length parameter to 50 (i.e., number of tokens to be generated), the temperature parameter and top_p (i.e., the creativity of the output) to 0.9 and 0.95 respectively, samples (i.e., total number of text outputs) to 100, and batch size (i.e., number of text outputs generated at a time) to 20. As a result, the language model has produced 2000 items per category. These items have been screened by a researcher, and the items deemed as highly relevant and clear. They have been added to the item pool. This has led to additional 73 items, with 29 referring to the content, 16 referring to the person in the video, and 28 referring to other aspects.

Synthesis and the proposed factorial structure

Altogether, the whole procedure has led to 419 items, with 90 items pertaining to the content, 121 items pertaining to the person in the video, and 208 items pertaining to other aspects.

While we have started off with three categories (or potential factors) - namely content, person in the video, and other aspects - established based on literature review (e.g. Hameleers, Van Der Meer and Dobber, 2023), the think-aloud techniques and the psychometric item generator has resulted in a large number of rather diverse items categorised into the "other" category. As such, we have additionally used inductive thematic analysis within this broad topic. Specifically, each item has been further coded into a more specific topic, such as "suspicious source", "low quality of the video", and "number of views, followers, subscribers". These specific topics have then been combined into three broader categories – source of the video, other video characteristics/technical aspects, and broader contextual characteristics. Once we have divided the items pertaining to other aspects into these three categories, we have ended up with 68 items referring to the source of the video, 82 items referring to technical aspects, and 58 items referring to the broader context.

As the number of items has been far too high for the next content validity step, we have reduced the number of items by only keeping the items that were unique (i.e., different ways of developing items led to items that severely overlapped) and general enough (i.e., some items were so specific that they would only be appropriate for responding to videos with certain content). The resulting first version of the questionnaire consisted of 123 items (content: 29 items, person in the video: 36 items, source of the video: 19 items, other video characteristics: 29 items, contextual characteristics: 10 items). This number has been further reduced in the next phases. Similarly, the number of items and the proposed factorial structure will be adjusted to empirical data if necessary.

Step 2: Shortening, adapting, and translating the scale

The first version of the questionnaire, developed in Step 1, was administered online to 13 experts (46.2% male, 38.5% female, 15.4% other or preferred not to disclose their gender; age: M = 36.4 years, SD = 14.05), working in various fields, such as psychology, philosophy, media studies, and computer science.

Participants were first informed about the study's aims, their rights as participants, and the scale we have been developing. After consenting to participate in the study, participants were asked to read each item, assess its relevance and clarity, and provide any additional comments (such as suggestions regarding the alternative wording). The items were presented in five blocks that correspond with the proposed dimensions of the questionnaire – 1) content of the video, 2) person in the video, 3) source of the video, 4) other video characteristics, and 5) contextual characteristics, with each block being accompanied by a short description (e.g., "The first group of items refers to video's content (i.e., the message conveyed in the video)"). In the case of relevance, participants were asked to assign each item one of the following values: 1 (the item is essential), 2 (the item is useful but not necessarily essential), or 3 (the item is not needed). Similarly, in the case of clarity, they were asked to assign each item one of the following values: 1 (the item is clear), 2 (the item should be partially modified to achieve the desired clarity), 3 (the item should be substantially modified to achieve the desired clarity), or 4 (the item is completely unclear). On average, participants needed about 20-30 minutes to complete the study.

During the analyses, we have calculated the content validity ratio (*CVR*; Lawshe, 1975), with higher values indicating higher relevance of the items (e.g., when the number of experts choosing "essential" is more than half, the *CVR* is somewhere between 0 and 1) and the content validity index (*CVI*; Rubio *et al.*, 2003), with higher values indicating higher clarity of the items (e.g., when the number of

experts indicating that the item is clear or only needs a partial modification is more than half, the *CVI* is somewhere between .50 and 1). For *CVR*, a cutoff of 8 (in the case of 11 experts) or 9 participants (in the case of 12 or 13 experts) indicating "essential" was used as the threshold (Wilson, Pan and Schumsky, 2012). For *CVI*, a conventional cutoff of .80 was used as the threshold. Additionally, we summarized participants' comments and used them to modify the items.

The results have showed that 31 items (8 related to the content of the video, 9 related to the person in the video, 7 related to the source of the video, and 7 related to technical aspects) and they have passed both the *CVR* and *CVI* acceptance threshold. Since none of the items from the "contextual characteristics" group has passed the *CVR* threshold, this group of items has been excluded entirely. Moreover, 5 of the 31 items have been slightly reworded according to participants' comments. Detailed results pertaining to the content validity of the retained items are presented in Figure 5 below.

Item	*CVR*	*CVI*	New version (according to comments)
I. CONTENT OF THE VIDEO			
The presented information seemed convincing.	.85 (12/13 ess.)	.92	/
The presented information seemed credible.	.69 (11/13 ess.)	.92	/
The message was not something I would expect from the person in the video.	.38 (9/13 ess.)	.92	The presented information was not something I would expect from the person in the video.
The message was consistent with my previous knowledge and information.	.54 (10/13 ess.)	.92	The presented information was consistent with my previous knowledge.
The presented information seemed plausible.	.54 (10/13 ess.)	1.00	/
The message conveyed something that I already know to be true.	.38 (9/13 ess.)	1.00	The presented information was something that I already know to be true.
The presented information seemed questionable.	.38 (9/13 ess.)	.92	/
The message contained errors.	.38 (9/13 ess.)	.92	The presented information contained errors.
II. PERSON IN THE VIDEO			
The face of the person in the video (or parts of it) was distorted.	.50 (9/12 ess.)	.92	/
The facial features of the person in the video changed during the video.	.50 (9/12 ess.)	.83	/
I found the voice of the person in the video to be different from their usual voice.	.67 (10/12 ess.)	.92	/
The mouth movements of the person in the video did not completely match the sound.	.67 (10/12 ess.)	1.00	/
The face of the person in the video (or parts of it) was blurry.	.50 (9/12 ess.)	.92	/
I found the voice of the person in the video unnatural.	.50 (9/12 ess.)	.92	/
The person in the video behaved authentically.	.83 (11/12 ess.)	.83	/
The mouth of the person in the video was moving strangely.	.45 (8/11 ess.)	1.00	/
The gestures displayed by the person in the video did not seem natural.	.45 (8/11 ess.)	.82	The person's gestures in the video did not seem natural.
III. SOURCE OF THE VIDEO			
I am not familiar with the source that published the video.	.50 (9/12 ess.)	.92	/
It is unclear to me who is the original source of the video.	.50 (9/12 ess.)	.92	/
The source of the video is verified in some way.	.50 (9/12 ess.)	.92	/
The source of the video is well-known.	.67 (10/12 ess.)	1.00	/
The content of the video is consistent with what this source has published previously.	.50 (9/12 ess.)	.92	/
The source of the video seems credible.	.67 (10/12 ess.)	.83	/
The video was posted by a reputable source.	.67 (10/12 ess.)	.92	/
IV. OTHER VIDEO CHARACTERISTICS			
The background in the video contained irrelevant or out-of-place objects.	.50 (9/12 ess.)	.83	/
The background in the video seemed authentic.	.83 (11/12 ess.)	1.00	/
I noticed a lot of flickering and lighting anomalies throughout the video.	.67 (10/12 ess.)	.83	/
Something about this video seemed off.	.50 (9/12 ess.)	.92	/
The video was visibly edited.	.50 (9/12 ess.)	.83	/
The audio was low quality.	.50 (9/12 ess.)	.92	/
The video quality was inconsistent.	.50 (9/12 ess.)	1.00	/

Figure 5. Content validity of the retained items.

Based on the content validity results, we created the second version of the questionnaire (together with instructions and the proposed scoring key), which can be seen in Figures 6 and 7 below.

Perceived video trustworthiness

The following questionnaire contains items that aim to capture your perception of the video you just watched (i.e., your opinion about the content, source, the person in the video, and technical aspects). Please read each item carefully and indicate your agreement using a 7-point scale ranging from »*Strongly disagree*« to »*Strongly agree*«. If you feel that you cannot answer a particular item, please choose »*Neutral*«.

		Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
1.	I noticed a lot of flickering and lighting anomalies throughout the video.	1	2	3	4	5	6	7
2.	The presented information seemed convincing.	1	2	3	4	5	6	7
3.	The video was visibly edited.	1	2	3	4	5	6	7
4.	The mouth movements of the person in the video did not completely match the sound.	1	2	3	4	5	6	7
5.	The background in the video contained irrelevant or out-of-place objects.	1	2	3	4	5	6	7
6.	The presented information seemed plausible.	1	2	3	4	5	6	7
7.	I found the voice of the person in the video unnatural.	1	2	3	4	5	6	7
8.	The person in the video behaved authentically.	1	2	3	4	5	6	7
9.	I found the voice of the person in the video to be different from their usual voice.	1	2	3	4	5	6	7
10.	The audio was low quality.	1	2	3	4	5	6	7
11.	The presented information was something that I already know to be true.	1	2	3	4	5	6	7
12.	Something about this video seemed off.	1	2	3	4	5	6	7
13.	The background in the video seemed authentic.	1	2	3	4	5	6	7
14.	The source of the video is verified in some way.	1	2	3	4	5	6	7
15.	The facial features of the person in the video changed during the video.	1	2	3	4	5	6	7
16.	The presented information was not something I would expect from the person in the video.	1	2	3	4	5	6	7
17.	The person's gestures in the video did not seem natural.	1	2	3	4	5	6	7

Figure 6. Second version of the questionnaire (part 1).

18.	The video quality was inconsistent.	1	2	3	4	5	6	7
19.	The source of the video is well-known.	1	2	3	4	5	6	7
20.	The face of the person in the video (or parts of it) was distorted.	1	2	3	4	5	6	7
21.	The presented information was consistent with my previous knowledge.	1	2	3	4	5	6	7
22.	The source of the video seems credible.	1	2	3	4	5	6	7
23.	The mouth of the person in the video was moving strangely.	1	2	3	4	5	6	7
24.	The presented information seemed questionable.	1	2	3	4	5	6	7
25.	The face of the person in the video (or parts of it) was blurry.	1	2	3	4	5	6	7
26.	I am not familiar with the source that published the video.	1	2	3	4	5	6	7
27.	The presented information contained errors.	1	2	3	4	5	6	7
28.	The content of the video is consistent with what this source has published previously.	1	2	3	4	5	6	7
29.	The video was posted by a reputable source.	1	2	3	4	5	6	7
30.	The presented information seemed credible.	1	2	3	4	5	6	7
31.	It is unclear to me who is the original source of the video.	1	2	3	4	5	6	7

Internal scoring key

Trustworthiness of the content of the video: (I2 + I6 + I11 + I16R + I21 + I24R + I27R + I30) / 8

Trustworthiness of the person in the video: (I4R + I7R + I8 + I9R + I15R + I17R + I20R + I23R + I25R) / 9

Trustworthiness of the source of the video: (I14 + I19 + I22 + I26R + I28 + I29 + I31R) / 7

Trustworthiness of the technical features: (I1R + I3R + I5R + I10R + I12R + I13 + I18R) / 7

Figure 7. Second version of the questionnaire (part 2).

We have also prepared the Slovene and Italian version of the questionnaire using the translation-backtranslation procedure. Specifically, a researcher who is bilingual in English and the target language at a C1 or C2 level of language competence has translated the questionnaire to the target language, another researcher (with equivalent language competence) has translated the questionnaire back to English, and both researchers together have resolved the inconsistencies and finalized the translation. The resulting translations can be seen in Figure 8 below.

English version	Slovene version	Italian version
Perceived video trustworthiness	Zaznana zaupljivost videoposnetka	Affidabilità percepita dei video
The following questionnaire contains items that aim to capture your perception of the video you just watched (i.e., your opinion about the content, source, the person in the video, and technical aspects). Please read each item carefully and indicate your agreement using a 7-point scale ranging from »Strongly disagree« to »Strongly agree«. If you feel that you cannot answer a particular item, please choose »Neutral«.	Spodnji vprašalnik vsebuje trditve, ki se nanašajo na vaše zaznavanje ravnokar ogledanega videoposnetka (tj. vaše mnenje o vsebini, viru, osebi na videoposnetku in tehničnih vidikih). Pozorno preberite vsako trditev in označite svoje strinjanje s pomočjo sedemstopenjske lestvice od "Sploh se ne strinjam" do "Povsem se strinjam". Če menite, da na določeno trditev ne morete odgovoriti, prosimo izberite odgovor "Nevtralno".	Il seguente questionario contiene item finalizzati a raccogliere informazioni sulla Sua percezione del video appena visto (ad esempio la Sua opinione sul contenuto, sulla fonte, sulla persona nel video e sugli aspetti tecnici). Per favore legga attentamente ogni item e indichi il suo grado di accordo usando la scala a 7 punti da "Fortemente in disaccordo" a "Fortemente d'accordo". Se ritiene di non poter rispondere a un item, per favore scelga "Neutro".
Strongly disagree; Disagree; Somewhat disagree; Neutral; Somewhat agree; Agree; Strongly agree	Sploh se ne strinjam; Ne strinjam se; Nekoliko se ne strinjam; Nevtralno; Nekoliko se strinjam; Strinjam se; Povsem se strinjam	Fortemente in disaccordo; In disaccordo; Un po' in disaccordo; Neutro; Un po' d'accordo; D'accordo; Fortemente d'accordo
I noticed a lot of flickering and lighting anomalies throughout the video.	Med ogledom videoposnetka sem opazil_a veliko utripanja in svetlobnih nepravilnosti.	Nel corso del video ho notato molti sfarfallii e anomalie di illuminazione.
The presented information seemed convincing.	Predstavljene informacije so se zdele prepričljive.	Le informazioni presentate sembravano convincenti.
The video was visibly edited.	Videoposnetek je bil vidno zmontiran.	Il video è visibilmente editato.
The mouth movements of the person in the video did not completely match the sound.	Premiki ust osebe na videoposnetku se niso popolnoma ujemali z zvokom.	I movimenti della bocca della persona nel video non corrispondevano completamente al suono.
The background in the video contained irrelevant or out-of-place objects.	Ozadje na videoposnetku je vsebovalo nerelevantne predmete ali predmete, za katere se zdi, da tja ne sodijo.	Lo sfondo del video presentava oggetti irrilevanti o fuori posto.
The presented information seemed plausible.	Predstavljene informacije so se zdele verjetne.	Le informazioni presentate sembravano plausibili.
I found the voice of the person in the video unnatural.	Glas osebe na videoposnetku se mi je zdel nenaraven.	La voce della persona nel video mi è sembrata innaturale.
The person in the video behaved authentically.	Oseba na videoposnetku se je vedla pristno.	La persona nel video si è comportata in modo autentico.
I found the voice of the person in the video to be different from their usual voice.	Imel_a sem občutek, da je bil glas osebe na videoposnetku drugačen od njenega običajnega glasu.	La voce della persona nel video mi è sembrata diversa dal solito.
The audio was low quality.	Zvok je bil slabe kakovosti.	La qualità dell'audio era bassa.
The presented information was something that I already know to be true.	Predstavljene informacije so bile nekaj, za kar že vem, da drži.	Sapevo già che le informazioni presentate erano vere.
Something about this video seemed off.	Nekaj na tem videoposnetku se je zdelo nenavadno.	Nel video c'era qualcosa che non mi tornava.
The background in the video seemed authentic.	Ozadje na videoposnetku se je zdelo pristno.	Lo sfondo del video sembrava autentico.
The source of the video is verified in some way.	Vir videoposnetka je bil na nek način preverjen.	La fonte del video è in qualche modo verificata.
The facial features of the person in the video changed during the video.	Obrazne značilnosti osebe na videoposnetku so se med videoposnetkom spreminjale.	Le caratteristiche facciali della persona nel video sono cambiate nel corso del video.
The presented information was not something I would expect from the person in the video.	Predstavljene informacije niso bile skladne s tem, kar bi pričakoval_a od osebe na videoposnetku.	Dalla persona presentata nel video non mi sarei aspettata/o le informazioni date.
The person's gestures in the video did not seem natural.	Geste osebe na videoposnetku se niso zdele naravne.	La gestualità della persona nel video non sembrava naturale.
The video quality was inconsistent.	Kakovost videoposnetka je bila nedosledna.	La qualità del video era incoerente.

The source of the video is well-known.	Vir videoposnetka je splošno poznan.	La fonte del video è nota.
The face of the person in the video (or parts of it) was distorted.	Obraz osebe na videoposnetku (ali njegovi deli) je bil popačen.	La faccia della persona nel video (o parti di essa) era distorta.
The presented information was consistent with my previous knowledge.	Predstavljene informacije so bile skladne z mojim predhodnim znanjem.	Le informazioni presentate erano coerenti con la mia conoscenza precedente.
The source of the video seems credible.	Vir videoposnetka se zdi verodostojen.	La fonte del video sembra credibile.
The mouth of the person in the video was moving strangely.	Usta osebe na videoposnetku so se čudno premikala.	La bocca della persona nel video si muoveva in modo strano.
The presented information seemed questionable.	Predstavljene informacije so se zdele vprašljive.	Le informazioni presentate sembravano discutibili.
The face of the person in the video (or parts of it) was blurry.	Obraz osebe na videoposnetku (ali njegovi deli) je bil zamegljen.	La faccia della persona nel video (o parti di essa) era sfocata.
I am not familiar with the source that published the video.	Ne poznam vira, ki je objavil videoposnetek.	Non ho familiarità con la fonte che ha pubblicato il video.
The presented information contained errors.	Predstavljene informacije so vsebovale napake.	Le informazioni presentate contenevano errori.
The content of the video is consistent with what this source has published previously.	Vsebina videoposnetka se ujema z vsebinami, ki jih je ta vir objavljal že prej.	Il contenuto del video è coerente con quanto questa fonte ha pubblicato precedentemente.
The video was posted by a reputable source.	Videoposnetek je objavil ugleden vir.	Il video è stato pubblicato da una fonte rispettabile.
The presented information seemed credible.	Predstavljene informacije so se zdele verodostojne.	Le informazioni presentate sembravano credibili.
It is unclear to me who is the original source of the video.	Nejasno je, kdo je izvorni vir videoposnetka.	Non mi è chiaro quale sia la fonte originale del video.

Figure 8. Translations to Slovene and Italian.

Looking ahead: Step 3 (Evaluation of psychometric characteristics)

As written in the SOLARIS project proposal, we have initially intended to test the questionnaire for the first time in our Use Case 1. However, to ensure quality measurement of perceived trustworthiness in the next stages of the project, we will psychometrically evaluate the questionnaire prior to the first use case. Specifically, we will conduct an online cross-sectional study in United States, Slovenia, and Italy (N ~ 300 per country) and perform analyses related to factorial structure, including invariance testing, internal consistency (Cronbach's alpha), and convergent/discriminant validity.

4 Generative AI for good

Despite the fact that generative artificial intelligence has been used in harmful ways and has caused negative effects, as laid out in the previous chapters, it also has the potential to be used for good purposes. Scientific advancements and progress in artificial intelligence technologies offer opportunities to create tools and apply them for social good to tackle the challenges of today's fast-paced society (Tomašev *et al.*, 2020) and Solaris addressed this issue too.

4.1. Artificial intelligence for social good

With the increasing development and employment of artificial intelligence across multiple scientific disciplines, a relatively new research field has emerged: Artificial intelligence for social good (AI4SG). AI4SG focuses on the utilizing artificial intelligence technologies to create positive effects in line with the fulfillment of the United Nations Sustainable Development Goals. In order to do so and increase effectiveness, AI4SG focuses more on applications rather than on textual output (Stewart, 2021).

Based on the following working definition, Floridi et al. (2020)created a model with seven essential factors to design AI for social good: "AI4SG = the design, development, and deployment of AI systems in ways that (i) prevent, mitigate or resolve problems adversely affecting human life and/or the wellbeing of the natural world, and/or (ii) enable socially preferable and/or environmentally sustainable developments." (Floridi *et al.*, 2020, pp. 1773–1774).

Factors	Corresponding best practices	Corresponding ethical principl
Falsifiability and incremental deployment	Identify falsifiable requirements and test them in incremental steps from the lab to the "outside world"	Nonmaleficence
Safeguards against the manipulation of predictors	Adopt safeguards which (i) ensure that non-causal indicators do not inappropriately skew interventions, and (ii) limit, when appropriate, knowledge of how inputs affect outputs from AI4SG systems, to prevent manipulation	Nonmaleficence
Receiver-contextualised intervention	Build decision-making systems in consultation with users interacting with and impacted by these systems; with understanding of users' characteristics, the methods of coordination, the purposes and effects of an intervention; and with respect for users' right to ignore or modify interventions	Autonomy
Receiver-contextualised explanation and transparent purposes	Choose a Level of Abstraction for AI explanation that fulfils the desired explanatory purpose and is appropriate to the system and the receivers; then deploy arguments that are rationally and suitably persuasive for the receiver to deliver the explanation; and ensure that the goal (the system's purpose) for which an AI4SG system is developed and deployed is knowable to receivers of its outputs by default	Explicability
Privacy protection and data subject consent	Respect the threshold of consent established for the processing of datasets of personal data	Nonmaleficence; autonomy
Situational fairness	Remove from relevant datasets variables and proxies that are irrelevant to an outcome, except when their inclusion supports inclusivity, safety, or other ethical imperatives	Justice
Human-friendly semanticisation	Do not hinder the ability for people to semanticise (that is, to give meaning to, and make sense of) something	Autonomy

Figure 9 9. Floridi's model.

The following graphic depicts the guidelines for future AI4SG projects, presented by Tomašev et al.*(Tomašev* et al.*, 2020, p. 33)* in order to facilitate successful collaborations across different types of organizations aiming to utilise artificial intelligence for sustainable development:

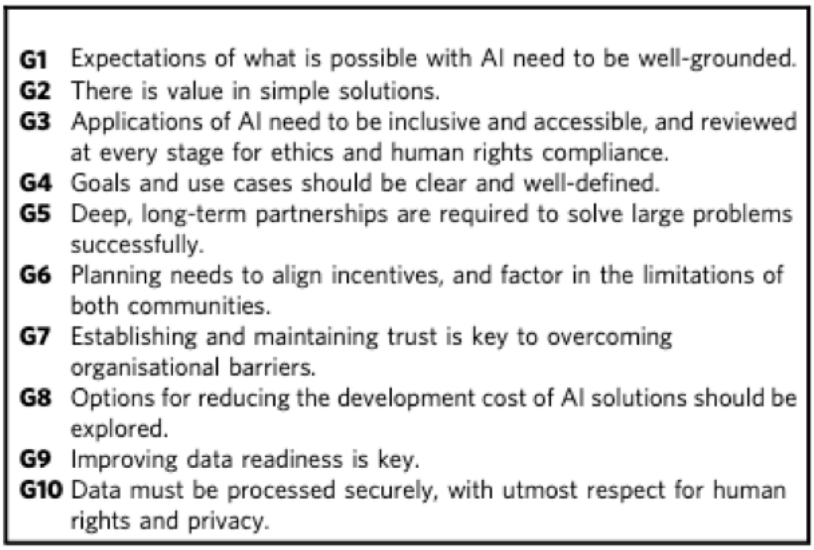

Figure 10. Guidelines for future AI4SG Initiatives.

These factors and guidelines can aid as indicators when designing artificial intelligence for good and conceptualising projects to create a positive social impact.

4.2. Social issues that artificial intelligence could help address.

Artificial intelligence will likely be used to address the challenges of our present and future. The possible fields and sectors of applying these technologies are vast. Thus, only a small fraction can be touched upon for this chapter.

Hager et al., with support from the Computing Community Consortium (CCC), identified several social issues that artificial intelligence could help address, including: justice, economic development, workforce development, public safety, policing and education (Hager *et al.*, 2019).

Artificial intelligence can also be used to tackle the challenges of climate change. Rolnick et al. suggest that machine learning and artificial intelligence could be applied to combat climate change, among others in the areas of electricity systems, transportation, industry, climate prediction and solar geoengineering (Rolnick *et al.*, 2019).

A few examples of how generative artificial intelligence has been used in ways to create a social benefit and bring a positive effect to individuals and society at large:

- The *Euphonia* project utilizes automatic speech recognition technologies and a machine learning algorithm to translate speech from individuals with a speech impediment in order to facilitate communication. Find out more: https://sites.research.google/euphonia/about/

- Artificial intelligence is used to detect fake news. Artificial intelligence evaluates the vast amounts of data and is thus able to detect fake news through pattern recognition (Ariwala, 2022).

- In the efforts to achieve global food security, artificial intelligence can play an important role, for example when it comes to crop management systems, water management as well as pest and disease management in crops (Chamara *et al.*, 2020).

Considering the deep fake technology, examples of positive uses in various fields can also be found.

The positive use of deep fakes can be observed in fields such as the entertainment sector, educational media, games, healthcare, fashion and e-commerce. Examples include the reduction of language barriers through speech translation, virtual fittings, stimulate people suffering from Alzheimer's through interacting with images of younger faces and the exploration of using GAN technology for medical advancements (Westerlund, 2019).

Deep fakes can also be employed to create an individual digital replica of a person, which could be used for long distance relationships or families. An example of such a deepfake could be a digital storytelling books, with the grandparents as readers. Despite this positive use, ethical concerns and psychological discomfort to be considered (Caporusso, 2021).

In the documentary 2020 *Welcome to Chechnya*, deep fakes were employed to protect activists' identities. Artificial intelligence was used to swap the real faces of the protagonists with faces of volunteers to ensure the activists' anonymity (Heilweil, 2020).

Leong et al. state several potential future positive applications of deep fakes, including:

- Creating artificial intelligence personas to increase motivation and attention in the field of education.
- Protecting personal identities through deep fake imagery.
- Supporting patients through artificial intelligence mental health counselors or virtual doctors.
- Employing artificial intelligence in the arts, storytelling and culture to tell the stories of the past and preserve (personal) world history, for example through interactive exhibitions (Danry *et al.*, 2022).

The positive examples of AI for good are manifold and will likely increase even further in the future. As technology evolves and is applied to our daily activities and interactions, it becomes ever more relevant to consider ethical and philosophical considerations and put relevant policies, frameworks and guidelines in place.

Artificial intelligence for good: partnerships and community

An important competent to ensuring that artificial intelligence benefits society in a positive way is the inclusion of all relevant key stakeholders as well as communities and citizens. Tomašev et al. encourage partnerships between key stakeholders such as researchers, academia and non-profit organisations to ensure that artificial intelligence can be leveraged to translate theory into practical social good and imitate an open societal discourse (2020).
A partnership between communities affected by the impact of artificial intelligence as well as the designers of such technologies is desirable in order to facilitate inclusive design and lasting benefit while also taking into account the unlikeliness that the impact of artificial intelligence effects can be exclusively positive (Bondi *et al.*, 2021).
The SOLARIS project considers these considerations by employing value-sensitive design approaches and a co-creative methodology as well as a transdisciplinary approach. The project will involve citizens in co-creation activities, crowdsourcing and participatory consultations to enquire what values citizens what GANs' to be designed with, what social issues should be addressed through creative engagement made by GANs, and to assess citizens' and stakeholders' expectations on GANs. (See SOLARIS DoA for additional information).

4.3. 'Soft fakes' as AI for good

In this section, we add some examples concerning the contents of section 2.3. above. According to Citron and Chesney (2018), videos that are partially fake but contain some truth can have many applications in social fields and discussions. In education, we can create historical documentaries that engage students and illustrate events by reconstructing the 'real' face of Julius Caesar using images of statues depicting him. Instead of deceased personalities, we could use their voices to convey positive messages, such as Sandro Pertini, former President of the Republic in Italy, or Mahatma Gandhi in India.
In the political arena, it is possible to enable candidates who lack the time to record specific content physically or to address voters, or those who do not know the language, such as migrants with voting rights, to have their digital alter egos do it for them. This way, their speeches and values can be appropriated and spread effectively.

Citron and Chesney discuss the use of Generative AIs in film special effects, where it has been possible to recreate deceased actors' computer images. However, this technology can also be used for thanatological purposes, such as creating videos of deceased loved ones to allow them to remain in contact with those still alive (Van Doorn *et al.*, 2021).

These videos would be different from deep fakes and considered 'soft fakes' (Santangelo, 2022). This technology can be explored to promote innovative and progressive uses of Generative AIs. It can be used to communicate verified historical notions, messages of social utility and concepts that politicians want to share with their constituents. A relevant example of this technology being used for good is the HBO documentary film, already mentioned, *Welcome to Chechnya*, which depicts the persecution of LGBTQ+ people by the Chechen dictatorship. The identities of these individuals were masked by merging their faces with those of volunteers. This use of Generative AIs has created a new type of personal identity, created to protect people, which Floridi (2018) defines "ectype".

Another example of using Generative AIs for good is the *Malaria Must Die* campaign, which featured former English football star David Beckham. In this campaign, Beckham addressed the camera in nine different languages, and the voices of professional actors pronounced the campaign text in their native languages. The use of A.I. technology edited the lip movements of Beckham to make all nine languages appear equally authentic, making the message more engaging for speakers of Swahili or Portuguese.

In the current state of research, soft fakes are effective, unlike the cases of Lola Flores and Anthony Bourdain mentioned above (paragraph 2.3), where they lacked the following four characteristics:

(1) A clear and predominant public utility.
(2) A recognisable author, preferably institutions and not private citizens or companies.
(3) An explicit statement in the text that it was produced through generative A.I.
(4) Controlled production and dissemination of the text.

Therefore, the necessary or optional relationships between these four characteristics must be better understood to prevent viral dissemination effects, which could lead to new forgeries.

References

Abbas, M. (2023) 'Socio-Technical Theory: A review', in S. Papagiannidis (ed.) *TheoryHub Book*. Available at: https://open.ncl.ac.uk /.

Ariwala, P. (2022) *Is artificial intelligence the key to combat fake news?* Available at: https://marutitech.com/artificial-intelligence-fake-news/ (Accessed: 3 May 2023).

Bassano, C., Chessa, M. and Solari, F. (2023) 'Visual working memory in immersive visualization: a change detection experiment and an image-computable model', *Virtual Reality*, 27(3), pp. 2493–2507. Available at: https://doi.org/10.1007/s10055-023-00822-y.

Bondi, E. *et al.* (2021) 'Envisioning Communities: A Participatory Approach Towards AI for Social Good', in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society*, Virtual Event USA: ACM, pp. 425–436. Available at: https://doi.org/10.1145/3461702.3462612.

Caporusso, N. (2021) 'Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology', in T. Ahram (ed.) *Advances in Artificial Intelligence, Software and Systems Engineering*. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), pp. 235–241. Available at: https://doi.org/10.1007/978-3-030-51328-3_33.

Chamara, R.M.S.R. *et al.* (2020) 'Role of artificial intelligence in achieving global food security: a promising technology for future', *Sri Lanka Journal of Food and Agriculture*, 6(2), pp. 43–70. Available at: https://doi.org/10.4038/sljfa.v6i2.88.

Cattlau, J. 2020. "Project Euphonia's new step: 1,000 hours of speech recordings" https://blog.google/outreach-initiatives/accessibility/project-euphonia-1000-hours-speech-recordings/. Last accessed on 2023-05-03.

Ceriani, G., Marketing Moving. L'approccio semiotico, Milano, FrancoAngeli, 2001.

Chesney, R. and Citron, D.K. (2018) 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security', *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3213954.

Danry, V. *et al.* (2022) 'AI-Generated Characters: Putting Deepfakes to Good Use', in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. *CHI '22: CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, pp. 1–5. Available at: https://doi.org/10.1145/3491101.3503736.

Eco, U. (1976) *A theory of semiotics*. Bloomington: Indiana University Press (Advances in semiotics).

Eco, U. (1991) *The limits of interpretation*. Bloomington: Indiana Univ. Press (Advances in semiotics).

Floch, J.-M. (2000) *Visual identities*. London New York: Continuum (Continuum studies in semiotics).

Floridi, L. (2014) *The fourth revolution how the infosphere is reshaping human reality*. Oxford: Oxford university press.

Floridi, L. (2018) 'Artificial Intelligence, Deepfakes and a Future of Ectypes', *Philosophy & Technology*, 31(3), pp. 317–321. Available at: https://doi.org/10.1007/s13347-018-0325-3.

Floridi, L. *et al.* (2020) 'How to Design AI for Social Good: Seven Essential Factors', *Science and Engineering Ethics*, 26(3), pp. 1771–1796. Available at: https://doi.org/10.1007/s11948-020-00213-5.

Geninasca, J. and Sadoulet, P. (2000) *La parole littéraire*. Paris: Presses universitaires de France (Formes sémiotiques).

Götz, F.M. *et al.* (2023) 'Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development.', *Psychological Methods* [Preprint]. Available at: https://doi.org/10.1037/met0000540.

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

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

Hager, G.D. *et al.* (2019) 'Artificial Intelligence for Social Good'. Available at: https://doi.org/10.48550/ARXIV.1901.05406.

Hameleers, M., Van Der Meer, T.G.L.A. and Dobber, T. (2022) 'You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media', *Social Media + Society*, 8(3), p. 205630512211163. Available at: https://doi.org/10.1177/20563051221116346.

Hameleers, M., Van Der Meer, T.G.L.A. and Dobber, T. (2023) 'They Would Never Say Anything Like This! Reasons To Doubt Political Deepfakes', *European Journal of Communication*, p. 02673231231184703. Available at: https://doi.org/10.1177/02673231231184703.

Heilweil, R. (2020) *How deepfakes could actually do some good*. Available at: https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous- artifcial-intelligence-welcome-to-chechnya (Accessed: 3 May 2023).

Hwang, Y., Ryu, J.Y. and Jeong, S.-H. (2021) 'Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education', *Cyberpsychology, Behavior, and Social Networking*, 24(3), pp. 188–193. Available at: https://doi.org/10.1089/cyber.2020.0174.

Ihde, D. (1990) *Technology and the lifeworld: from garden to earth*. Bloomington: Indiana University Press (The Indiana series in the philosophy of technology).

Latour, B. (1987) *Science in action: how to follow scientists and engineers through society*. Cambridge, Mass: Harvard University Press.

Lawshe, C.H. (1975) 'A QUANTITATIVE APPROACH TO CONTENT VALIDITY [1]', *Personnel Psychology*, 28(4), pp. 563–575. Available at: https://doi.org/10.1111/j.1744-6570.1975.tb01393.x.

Lee, J. and Shin, S.Y. (2022) 'Something that They Never Said: Multimodal Disinformation and Source Vividness in Understanding the Power of AI-Enabled Deepfake News', *Media Psychology*, 25(4), pp. 531–546. Available at: https://doi.org/10.1080/15213269.2021.2007489.

Leong, J., Danry, V., 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." CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA.

Marrone, G. (2014) *The invention of the text*. Italy? Mimesis International.

Marrone, G. (2021) *Introduction to the Semiotics of the Text*. De Gruyter. Available at: https://doi.org/10.1515/9783110688986.

Ng, Y.-L. (2023) 'An error management approach to perceived fakeness of deepfakes: The moderating role of perceived deepfake targeted politicians' personality characteristics', *Current Psychology*, 42(29), pp. 25658–25669. Available at: https://doi.org/10.1007/s12144-022-03621-x.

Pinch, T. and Leuenberger, C. (2006) 'Researching Scientific Controversies: The S&TS Perspective', *EASTS Conference" Science Controversy and Democracy* [Preprint].

Polidoro, P. (2008) *Che cos'è la semiotica visiva*. 1. ed. Roma: Carocci (Le bussole Scienze della comunicazione, 314).

Raynaud, D. (2003) *Sociologie des controverses scientifiques*. Paris: Presses universitaires de France (Sociologies).

Reber, B. (2006) 'Les controverses scientifiques', in *La science au présent*. Parigi (Encyclopaedia Universalis), pp. 156–159.

Rolnick, D. *et al.* (2019) 'Tackling Climate Change with Machine Learning'. Available at: https://doi.org/10.48550/ARXIV.1906.05433.

Rubio, D.M. *et al.* (2003) 'Objectifying content validity: Conducting a content validity study in social work research', *Social Work Research*, 27(2), pp. 94–104. Available at: https://doi.org/10.1093/swr/27.2.94.

Russo, F. (2022) *Techno-scientific practices: an informational approach*. Lanham Boulder New York London: Rowman & Littlefield.

Santangelo, A. (2022) 'Il futuro del volto nell'era dei deep fake", Il metavolto', *FACETS Digital Press*. Edited by M. Leone, pp. 18–41.

Sharkey, A. and Sharkey, N. (2021) 'We need to talk about deception in social robotics!', *Ethics and Information Technology*, 23(3), pp. 309–316. Available at: https://doi.org/10.1007/s10676-020-09573-9.

Shin, S.Y. and Lee, J. (2022) 'The Effect of Deepfake Video on News Credibility and Corrective Influence of Cost-Based Knowledge about Deepfakes', *Digital Journalism*, 10(3), pp. 412–432. Available at: https://doi.org/10.1080/21670811.2022.2026797.

Sparrow, R. and Sparrow, L. (2006) 'In the hands of machines? The future of aged care', *Minds and Machines*, 16(2), pp. 141–161. Available at: https://doi.org/10.1007/s11023-006-9030-6.

Stewart, M. (2021) 'Introduction to AI for Social Good.', 4 August. Available at: https://medium.com/towards-data-science/introduction-to-ai-for-social-good-875a8260c60f (Accessed: 3 May 2023).

Tahir, R. *et al.* (2021) 'Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos', in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. *CHI '21: CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, pp. 1–16. Available at: https://doi.org/10.1145/3411764.3445699.

Thaw, N.N. *et al.* (2021) 'How Are Deepfake Videos Detected? An Initial User Study', in C. Stephanidis, M. Antona, and S. Ntoa (eds) *HCI International 2021 - Posters*. Cham: Springer International Publishing (Communications in Computer and Information Science), pp. 631–636. Available at: https://doi.org/10.1007/978-3-030-78635-9_80.

Tomašev, N. *et al.* (2020) 'AI for social good: unlocking the opportunity for positive impact', *Nature Communications*, 11(1), p. 2468. Available at: https://doi.org/10.1038/s41467-020-15871-z.

Van Doorn, M. *et al.* (2021) *Real Fake: Playing with Reality in the Age of AI, Deepfakes and the Metaverse*. Groningen: Ludibrium publishers.

Verbeek, P.-P. (2011) *Moralizing technology: understanding and designing the morality of things*. Chicago, Ill.: The Univ. of Chicago Press.

Westerlund, M. (2019) 'The Emergence of Deepfake Technology: A Review', *Technology Innovation Management Review*, 9(11), pp. 39–52. Available at: https://doi.org/10.22215/timreview/1282.

Wilson, F.R., Pan, W. and Schumsky, D.A. (2012) 'Recalculation of the Critical Values for Lawshe's Content Validity Ratio', *Measurement and Evaluation in Counseling and Development*, 45(3), pp. 197–210. Available at: https://doi.org/10.1177/0748175612440286.