Strengthening democratic engagement through value-based generative adversarial networks

D2.1 SOLARIS
Technical and socio-political reconnaissance of GANs-generated contents

Lead Author: UC3M
With contributions from: UvA, LUMSA, CINI, DEXAI, UM
Reviewers: ANSA, IMA

Deliverable nature:	R – Document, report
Dissemination level: (Confidentiality)	PU - Public
Delivery date:	September 2024
Version:	2
Total number of pages:	49
Keywords:	Generative AI; GANs; Deepfakes

Executive summary

This deliverable provides a general overview of the problems concerning AI-generated content from a multidisciplinary perspective.
Deliverable D2.2, due at month 10, will provide an in-depth analysis of several issues listed here.
The first chapter briefly defines the Adversarial Generative Network types developed from 2017 to 2023. A link between AI-generated content, deepfakes and Generative Adversary Networks (GANs) is also drawn here.
In the second chapter, the negative policy implications of these technologies are explored, particularly concerning their abuse or misuse and the biases they can foster.
The third chapter emphasises the risk that Generative AI poses at the level of international relations since GANs and AI have also started to create internal and external problems being utilised in electoral campaigns, damaging the opponent's reputation or exacerbating tensions where they already exist.
Chapter 4 explores the link between deepfakes and fake news from a semiotic perspective. Here, models are constructed to articulate different types of lies and fakes, discussed concerning deepfake images.
Chapter 5 is devoted to presenting a cross-legal approach to Generative AI and pointing out possible regulatory paths on a global scale.
In Chapter 6, the research team resolved to zoom into a particular case of distribution of deepfakes in the media: Bulgaria. The distribution of disinformation in this EU member creates an interesting use case in the media landscape. Therefore, in this chapter, we analyse how AI-generated content development has contributed to the spread of disinformation in the Bulgarian mediascape.

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**	2.1	**Title**	Technical and socio-political reconnaissance of GANs-generated contents
Work package	**Number**	2	**Title**	Definition & Mapping of GANs in social media
Task	**Number**	2.1	**Title**	Reconnaissance of GANs in social media

Date of delivery	**Contractual**	M8	**Actual**	M20
Status	version 2		☒Final version	
Nature	☒Report ☐Demonstrator ☐Other ☐ORDP (Open Research Data Pilot)			
Dissemination level	☒Public ☐Confidential			

Authors (partners)	UC3M - LUMSA			
Responsible author	**Name**	Piero Polidoro		
	Partner	LUMSA	**E-mail**	p.polidoro@lumsa.it

Summary (for dissemination)	Multidisciplinary overview of the problems and political risks of generative AI content
Keywords	Generative AI; GANs; Deepfakes

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
06-09-2024	2		Reviewer's requests for changes incorporated

Acknowledgement

Disclaimer

Table of Contents

List of Figures

Abbreviations and acronyms

AI	Artificial intelligence
DCGAN	Deep Convolutional GAN
GAN	Generative Adversary Networks
GLOBSEC	Global think tank
GPT	Generative Pre-trained Transformer
HAARP	High Frequency Active Auroral Research Program
IP	Intellectual property
NLG	Natural language generation
PGAN	Progressive GAN
WGAN	Wasserstein GAN

1 Technical definition of GANs and requirements

AI models can be categorised as discriminative and generative models. Discriminative models provide a function related to input data, while generative models generate new data. One example of a generative model is the Generative Adversarial Network (GAN), which consists of a discriminator and generator model. The discriminator distinguishes between real and fake input, while the generator creates new data. The goal is to create a robust generator to create realistic images that the discriminator cannot distinguish from real ones. The models play an adversarial game where the generator tries to create realistic images, and the discriminator tries to identify which images are real or fake. A GAN will produce excellent results when the generator fools the discriminator into identifying images.

1.1 Types of GAN's architecture

Vanilla GANs (2014) are GANs presented in the original paper by Ian Goodfellow (Goodfellow *et al.*, 2014). The GAN setup is the following. **G** represents the function that binds the noise **z** to the generated image, **X** is the real image, and **D** is the function accurately calculating the true content of the input image.

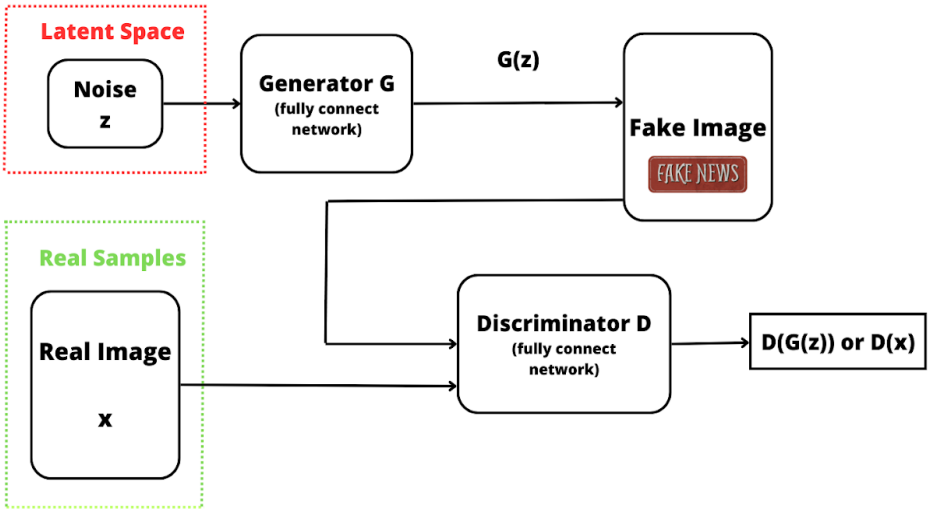


Figure 1

The generator architecture in Vanilla GAN consists of a series of dense layers, usually followed by nonlinear activation functions such as hyperbolic tangent (tanh) or ReLU.
The discriminator architecture in Vanilla GAN consists of a series of Dense layers followed by nonlinear activation functions such as ReLU or Leaky ReLU. The last layer of the discriminator uses a sigmoid activation function to produce a value between 0 and 1, representing the probability that the input image is real.
The training of a GAN takes place in two alternating phases: in the first phase, the discriminator learns to distinguish between real and false images while the generator remains unchanged; in the second phase, the generator tries to fool the discriminator, which in turn remains unchanged.
The results from Vanilla GAN are encouraging, but improvements are still needed.

Until now, GANs were unable to generate images with controllable features. This feature was introduced by **ConditionalGAN (cGANs)** in **2014** by Mirza (Mirza and Osindero, 2014). In other words, we can specify which class of examples we want to generate. The discriminator in a conditional GAN has the additional task of not only distinguishing between real and fake images, but also ensuring that the generated image satisfies the desired features. Both the generator and the discriminator in a CGAN have an additional parameter representing the desired class of the output among the available classes.

The **Deep Convolutional GAN (DCGAN)** is a model published in **2016** that improved the output of GANs by using transposed convolutions and batch normalization layers to generate high-resolution images. The generator takes a noise z as input and goes through a series of up-sampling, convolutional, and batch normalization layers. In an up-sampling layer, the input tensor is expanded to increase image resolution. DCGAN demonstrated GANs' ability to generate high-quality images. They include realistic images of people's faces, with fewer training cycles than Vanilla GAN[1]. DCGAN also introduced **vector arithmetic** in latent space (noise vector Z) to generate features not present in the domain, allowing for manipulation of the noise input to produce specific images. For example, manipulating latent vectors of a man with glasses, a man without glasses, and a woman without glasses can produce an image of a woman with glasses subtracting the first two and adding to the third.

Wasserstein GAN (WGAN) (2017) uses a simpler discriminator than traditional GANs and does not use the sigmoid activation function in the output layer of the discriminator. Instead, WGAN's discriminator only has an output that represents the evaluation of the Wasserstein distance function as a truth or falsity score. Using a function that returns the quality of the obtained image improves the stability in learning and the results obtained.

Progressive GAN (PGAN) (2017) is a variant of Generative Adversarial Network (GAN) in which image generation is divided into several progressive steps, starting from low resolution and gradually increasing the resolution to the desired resolution.
In PGAN, the goal has been to solve the problem of generating high-resolution images by separating the problem into smaller, easier-to-solve problems.
The system starts learning from a lower resolution , 4x4 to allow for simpler and more stable training and then increases the resolution step by step by introducing layers for the two models to get to high resolutions , for example 1024x1024 .The progressive training process allows the models to learn the basic features of low resolution images before moving to higher and more detailed resolutions, allowing the models to generate more realistic and consistent high resolution images.

CycleGAN (2017) is used for image transformation. In particular, this model can be used to convert images from one domain into images to another domain. For example, it could be used to convert images of horses to images of zebras. One of the main innovations of CycleGAN compared to other image transformation models is the ability to learn this transformation in both directions.
CycleGAN uses a technique called "cyclic consistency". In CycleGAN, two generators and two discriminators are used to ensure cyclic consistency. Specifically, the first generator converts source domain images to target domain images, while the second generator converts images generated by the first generator back to the original source domain.
The first discriminator evaluates the quality of the images generated by the first generator and distinguishes them from the real images of the target domain, while the second discriminator evaluates the quality of the images generated by the second generator and distinguishes them from the real images of the source domain.

DiscoGAN (2017) uses two generators: GAB and GBA and two discriminators DA and DB as in cycleGAN. The GAB network is trained to transform images from domain A to domain B, while the GBA network is trained to transform images from domain B to domain A. The two networks are trained simultaneously and alternately until the model is able to generate images that are consistent with both domains.
The difference between DiscoGAN and CycleGAN is that DiscoGAN uses two reconstruction loss, one for both the domains to account for the information of domain transformation whereas CycleGAN uses single cycle-consistency loss.

StyleGAN-1 (2019) combines Progressive GAN and neural style transfer. One of its key features is the progressive growth mechanism, which is similar to Progressive GAN. To generate an image, a constant 4x4x512 array is repeatedly passed through style blocks. However, unlike Progressive GAN, StyleGAN-1 uses a mapping network with 8 fully connected layers to generate a new vector w based on the latent space. The vector w is then transformed by a learned affine transformation (A) and sent to the synthesis network, which uses Adaptive Instance Normalization (AdaIN) to adapt the content and style of the synthesized image according to the provided style vector. The model also uses noise B before applying the AdaIN module to generate images consistent with a specific artistic style.To generate an image, a constant StyleGAN-2 (2020) and StyleGAN3 (2021) were introduced to address problems found in the previous version. StyleGAN-2 uses “residual connections” technology to solve the issue where some image features appear in line with the camera rather than the generated face. It also solves the “blob problem” by using a latent style vector. StyleGAN2-ADA improves image generation performance through a technology called “invertible data augmentation”, which results in better and more realistic images and allows training StyleGAN2 with limited data. StyleGAN3 corrects “texture sticking” to make transition animations more natural, allowing generated faces to rotate and translate smoothly.

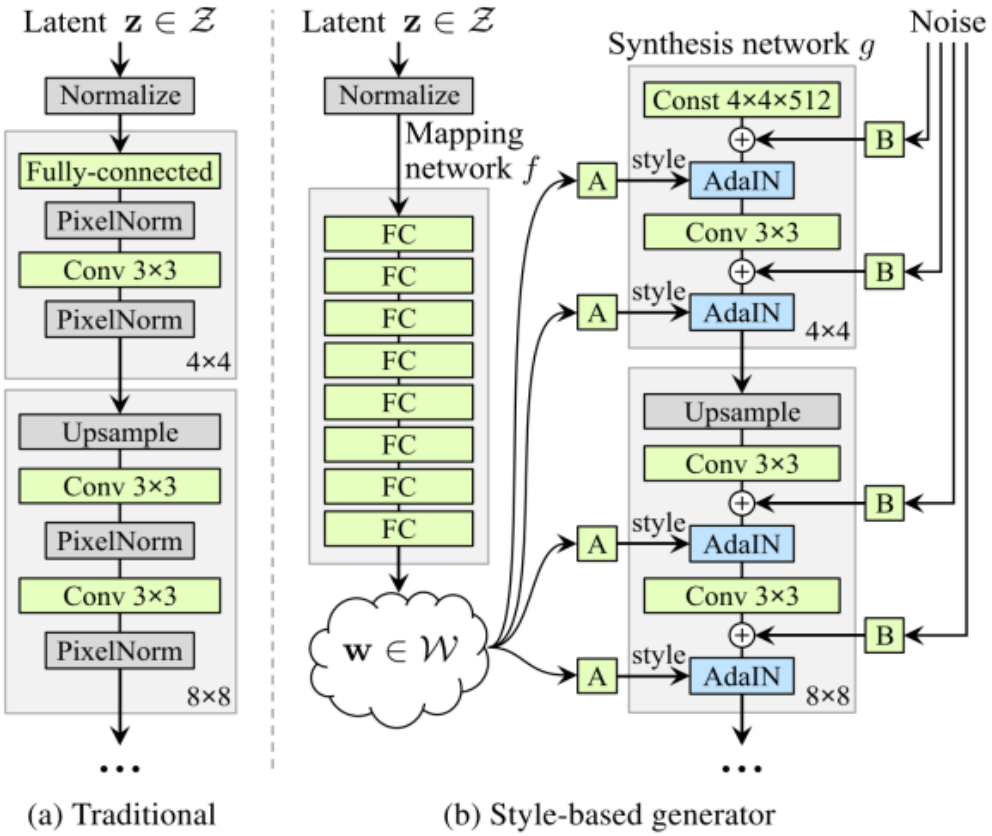


Figure 2

InsetGAN(2022) is one of the first GANs that generates plausible human body images. It uses multiple pre-trained GANs to improve and refine the details of the generated image. InsetGAN involves two main generators, both based on StyleGAN2. The first generator creates an image of a complete human body, including an initial face. The second generator generates a better quality face, which is then inserted into the previously generated body image through joint optimization. The optimization is done using latent space information to generate the final image in a way that minimizes an objective function that targets both the quality of the generated face and the consistency between face and body. GANs used to enhance other parts of the body are called "bugs" that are then inserted into the final model.

3D-Human GAN(2023) is an image generator trained using an adversarial generative neural network (GAN) designed to produce images of people in different positions and angles. The 3D-2D hybrid generator uses 3D geometric prioritization of the human body and image segmentation as input. The image generation process starts with selecting a pose, view, and appearance. This information is then passed to the GAN generator, which produces a synthetic image of a person in that specific pose and view. The generator is adapted to generate a 3D image from a 2D generator modulated by the 3D pose in input, resulting in better image quality than existing technologies.

1.2 From AI-generated content to deepfakes

Generative AI is a field within Artificial Intelligence that focuses on creating new data instances from the data it is trained on. This technique is used to generate real-life data instances, such as the 3D Avatars or Memojis commonly used in daily communication. The software used to create these avatars and videos produces such realistic images and videos that can easily be mistaken for real-life images.
In order to create new data instances, Generative AI models use available text, audio files, images, and videos. Those instruments are used to create a new dataset which is perfect and complete according to its own definition. The algorithms used in these models understand the patterns in the data that is fed to them and use that understanding to create a new version of the data.
Generative AI models are primarily Unsupervised AI models. These models learn about the distribution of the data that is fed to them and use that understanding to create new instances of data. Unlike Discriminative models, which are based on prediction based on conditional probabilities, Generative models work on finding the actual distribution of the dataset. They often use the Bayes theorem to predict the joint probability.
However, Generative models are computationally expensive, since they create new data instances. Discriminative models, on the other hand, discriminate between data instances and provide answers. For example, a Discriminative model would be able to determine whether an image represents a car or a bike, whereas a Generative model would be able to generate an image of a car based on given features. These models work on complex correlations and distributions to generate new data instances.

1.3 From GANs to deepfakes

Deepfakes are content generated using artificial intelligence, which appears realistic and authentic to humans. Generative Adversarial Networks (GANs) play a crucial role in deepfake technology[1]. There are different applications of deepfakes, including creative, productive, and unethical/malicious use

[1] We extend our sincere gratitude to one of the deliverable reviewers for bringing to our attention the rapidly evolving nature of deep fake technology. It is crucial for our project to stay updated with the latest advancements in this field, as they will impact capabilities and public perception. Specifically, deliverable D.6.1., scheduled for September 2025, is tasked with documenting the evolution of state-of-the-art technology from the beginning to the conclusion of the project.

cases. Some examples include recreating historical figures, movie translation, fashion, video game characters, stock images, and pornography. However, the use of deepfakes for impersonation or creating fake content can have severe consequences. In a typical deepfake setup, there is a source, target, and generated content. The source is the identity controlling the output, the target is the identity being faked, and the generated content is the result of transforming the source into the target.

Replacement techniques can be categorized into transfer and swap. Transfer involves transferring a specific aspect from one image to another, while swap involves exchanging two subjects in an image or video. In swap the replacement image maintains the characteristics of target image, hair, pose and so on.

Re-enactment methods are utilized to capture characteristics of the pose, expression, and gaze of the target.

- Gaze: Techniques in this area try to re-enact the generated output's gaze based on the source's eye movements/gaze. Useful for maintaining eye contact in videos.
- Mouth: in this case, the mouth movements of target are conditioned on the mouth movement of the source.
- Expression: these techniques are used to guide the expression of the destination based on the source
- Pose: the source drives the target for pose re-enactments.

Editing

Add, remove or alter certain aspects of the target entity to serve specific objectives (age, ethnicity, gender, hair), used in mobile app for being use case for fun (face app) or commercial use.

2 Negative Impacts of Deepfakes

2.1 GANs technology and democracy

Generative AI technologies are capable of producing high-quality inauthentic media content (e.g., images, video, audio, text) that is near-indistinguishable from authentic content. Such programs present a significant risk to democracy by directly interfering in politics, manipulating political discourse, and undermining democratic values of trust, human dignity, and equal participation.

In the upcoming discussions regarding the SOLARIS project, we will delve deeper into the concept of deep fakes. These are distinguished from simple image editing methods, such as those used in Photoshop. Before the rise of AI, video editing was primarily used to enhance images, for instance, in fashion and film photography. It was possible to alter documents and images, such as forging signatures or creating caricatures. However, with the advancement of AI-based techniques, we now differentiate between deep fakes and "cheapfakes", which are media altered by humans using accessible technologies with minimal time and effort. The distinction between '*shallowfakes*' as opposed to '*deepfakes*' is also used[2].

Another important point to note is that AI-based technologies allow for intervention in moving images, enabling the creation of modified video content. This means that deep fakes can achieve a high level of realism in various types of culturally reliable texts, such as documentaries, interviews, congressional statements, phone videos, and reels.

Political interference

The production of inauthentic audio-visual content depicting government officials, political figures, and influential media personalities doing or saying anything (so-called "deepfakes") could be used to undermine the reputation of these public figures, deceptively manipulate public opinion on political issues, disrupt international relations, and interfere with elections.

The current state of Generative AI programs is currently incapable of producing inauthentic audio-visual content of such a high quality as to be entirely undetectable and often require significant resources including large training datasets. However, this technology is rapidly developing in quality and will require less training data such that it will significantly increase the veracity of online disinformation and leave anyone vulnerable to targeting (Smith and Mansted, 2020).

High-profile disinformation

AI-generated media depicting high-profile public figures can be used to directly influence public perceptions of this figure, disrupt their established political relationships, and exploit their popularity to spread disinformation on various issues.
For example, at the beginning of the Russia-Ukraine war in March 2022, a video was circulated online depicting a low-quality deepfake of Ukrainian president Volodymyr Zelens'kyj. In the video, Zelens'kyj addresses the nation, stating that the Ukrainian military had failed, and that the public should surrender to Russian forces (Burgess, 2022).
In another example, in June 2023 several Russian radio and TV stations were hacked by an unknown group and broadcast a deepfake of Russian president Vladimir Putin. In the video, Putin announced the evacuation of civilians from three Russian regions on the border with Ukraine (*Il Post*, 2023).

[2] See, for instance, https://www.samsungsds.com/en/insights/what-are-cheapfakes.html. Last accessed on 27/08/2024. We want to thank the reviewer who suggested adding this clarification in the first part of the report.

Though AI-generated content can be of a very high technical quality as to appear indistinguishable to authentic content, it is unlikely that this form of high-profile disinformation will deceive the public. This is likely because such high-profile content is distributed through mainstream media channels and receives considerable scrutiny from fact-checkers in journalism, politics, and academia, as well as within general public discussion, as to be easily detected and debunked. However, this has yet to be thoroughly investigated and so SOLARIS will study the factors at play using a psychometric scale described in further detail in deliverable D2.2

Microfakes

Where AI-generated media featuring high-profile public figures is more likely to be debunked, deepfake content depicting figures and officials involved in small-scale politics, so-called "microfakes", may go undetected as such content is unlikely to be widely distributed and properly scrutinised (Ascott, 2020).
This form of small-scale disinformation featuring local politicians or officials addressing low-profile controversies (e.g., road quality, bypass development, cycle lanes) may appear convincing and interfere with local elections.
While the initial political impact of microfakes may be small, widespread production and distribution involving numerous targets represents a granular threat to democracy that can rapidly escalate and influence large-scale political issues.

Intimidation

AI-generated media could be used to intimidate, threaten, or otherwise harass individuals in order to influence their actions and statements or to deter political participation altogether. Notably, the production of deepfake pornographic content presents a significant risk to the reputations of public figures, particularly female public figures as the targets are almost exclusively women (Adjer *et al.*, 2019).
Even low-quality deepfake pornography that is noticeably fake risks reputational damage by objectifying and fetishising female public figures and enabling large-scale sexual harassment online (Van der Nagel, 2020). Such intimidation tactics may exacerbate existing equality issues such as gender representation in politics. As for these aspects, they are covered in Deliverable D.1.2., which has been delivered previously.
While high-profile examples of deepfake pornography predominantly feature women in the entertainment industry including actresses Emma Watson and Gal Gadot, with the arrival of deepfake apps such as DeepNude any social media user with a significant number of personal images on their public profile is a potential target (Harwell, 2018).

Exploiting credibility

Outside politics, Generative AI programs have been used to digitally recreate deceased actors and celebrities to either reprise a previous role or create an entirely new performance. These deepfake performances have not been explicitly authorized by the deceased but are legal with permission provided by their surviving relatives or by holders of their image rights.
While these deepfake performances are unlikely to be viewed as authentic, the use of trusted or admired public figures may lend credibility to or promote political sentiments or attitudes expressed in the content and that were not previously associated with the figure themselves.
For example, in 2019 it was announced that the upcoming Vietnam war film *Finding Jack* would feature a digitally recreated James Dean as an original character over sixty years on from his death in 1955 (Damiani, 2019). This deepfake performance associates Dean's iconic image with the political

issues of a conflict that occurred long after his death and this association could further help promote this particular representation of the war, as well as the ideological messages of the film's narrative.

Deepfake geography

Beyond replicating images of people, Generative AI programs can similarly be trained to generate images of landscapes and scenes, as exemplified in the production of AI-generated portraiture such as the *Portrait d'Edmond de Belamy* (Obvious, 2018). In the context of politics and disinformation, however, the production of deepfake satellite images presents risks to democracy and international relations.
Rather than deceiving the public and manipulating political sentiment, deepfake satellite images could be used to interfere with military and intelligence operations, for example, by misrepresenting the number of enemy personnel present at national borders or in disputed territory. Furthermore, such images could interfere with the activities of human rights investigators by hiding evidence of atrocities (Fingas, 2021).
The use of deepfake satellite images could deceptively incite distrust, animosity, and potentially conflict between nation states, especially those with historically strained political relations.

Speculative visualizations

AI image generators such as DALL-E are not limited to the individuals and can be used to visualize scenarios based on descriptive written statements. Such programs could be used to visualize hypothetical events that have not or have yet to, as well as events that have occurred but that have not been photographed or video recorded. For example, a 2023 art exhibition used an AI program to visualize the experiences of refugees detained in Australian offshore immigration processing centres on Nauru and Manus Island. Using the written statements of refugees as prompts, the AI program produced graphic deepfake photographs that emphasised the poor living conditions and violent occurrences within these centres (Doherty, 2023). While the above presents a positive example in that highlights the previously unknown suffering of refugees, this example also highlights the potential for Generative AI to visualize events from a particular political framing. As another, the US Republic party produced an advertisement attacking President Joe Biden which featured AI-generated images of hundreds of people crossing the US border, stating that Biden's re-election would lead to uncontrolled migration (Novak, 2023). Such speculative visualizations pose a risk to democracy and democratic values in that they could inauthentically manipulate public sentiment through misleading images.

Discourse manipulation

Beyond direct depictions of individuals, Generative AI could be used to produce various different forms of inauthentic content that can otherwise influence or manipulate public discourse around particular political issues.

Sock puppet profiles

Similarly to deepfake techniques, Generative AI programs can produce high-quality images of non-existent people which are being used to legitimise false actors operating online; so-called "sock puppet profiles". High-quality AI-generated images of people are freely available on websites such as ThisPersonDoesNotExist.com[3].

[3] https://this-person-does-not-exist.com/en - last accessed 17-07-2023

These images have been used as profile pictures for social media accounts to give the appearance of authenticity and allow anonymous operators to spread disinformation more easily (Smith and Mansted, 2020). If a significant number of sock puppet profiles operate in coordination, these profiles could be used to manipulate political debate and discussion online and present a particular risk to democracy. For example, in 2019 a LinkedIn profile for a research fellow at the Center for Strategic and International Studies at the University of Michigan named "Katie Jones" began connecting with US government and senate officials as part of a larger suspected espionage effort. Detection systems used to recognise deepfake images noted visual anomalies in Jones' profile picture that may indicate that it was AI-generated (Satter, 2019).

Artificial evidence

NLG programs are also capable of emulating the narrative styles of technical reports and scientific essays. Several researchers have begun listing ChatGPT as a co-author on papers and some scientific journals have allowed for this within their submission guidelines (Zheng and Zhan, 2023). For example, radiology researcher Som Biswas demonstrated the potential of NLG programs in scientific writing by publishing a paper written almost entirely by ChatGPT in the highly regarded *Radiology* journal (Biswas, 2023). As NLG programs lack contextual understanding, their outputs can include information that is fundamentally incorrect and distorted or provide falsified findings that are difficult for readers and reviewers to detect (Zheng and Zhan, 2023). In the context of democracy, NLG programs could be used to manufacture scientific literature and evidence supporting false concepts, bolstering the veracity of disinformation campaigns.

Automated text generation

Natural language generation (NLG) programs such as OpenAI's ChatGPT present one of the most significant risks to democracy and democratic values in that they excel at producing text that is both coherent and convincing in its argumentation with limited prompting or human intervention. In the hands of bad actors, language models may be misused to produce targeted disinformation on a massive scale.

Trained on curated datasets, these programs are capable of generating a range of diverse messages and narrative forms (e.g., articles, essays, social media posts) that logically argue specific false concepts and positions (e.g., climate change denial), copy the appealing narrative styles of conspiracy theories (e.g., QAnon), and express a particular worldview by employing certain language patterns such as the sexist use of "he" rather than "she" or "they" when referring to a generic subject (Buchanan *et al.*, 2021).

For example, OpenAI's GPT-3 language model has been used to generate text content promoting far-right extremist ideals that is not only coherent but demonstrably logical and convincing in its argumentation (McGuffie and Newhouse, 2020).

Undermining democratic values

Beyond the direct interventions discussed above, continued production and circulation of AI-generated media may have indirect impacts on democracy that could be significantly more damaging including exacerbating issues of mistrust, inappropriate deployment of AI technologies, and discouraging political participation.

Climate of uncertainty

While certainly the quality of deepfake content is rapidly increasing, media reports and public discourse on the topic often overexaggerate current technological capabilities and contribute to a climate of uncertainty in online media. Such a climate of uncertainty undermines trust in authentic content and previously trusted sources of information such as media organizations and government institutions (Vaccari and Chadwick, 2020). This presents a risk to democracy in that it could encourage cynicism and political disengagement, exacerbate issues of disinformation and radicalization, and incite violence and unrest.

For example, following the hospitalization and prolonged absence of Gabonese president Ali Bongo Ondimba from public life in 2018, the Gabonese government released a video address featuring the president to dispel rumours of his ill health or death. However, many speculated that the video address was a deepfake and, citing this as proof of deception, the Gabon military launched a failed coup against the government. The video's authenticity was later proven and Ondimba returned to public life in 2019 (Adjer *et al.*, 2019).

Epistemic shame

Democratic participation is reliant upon citizens feeling empowered and trusting their own capacity to form, reflect on, and deliberate their own political beliefs based on the information they receive. However, as AI-generated media becomes undetectable to the human eye, citizens may feel disempowered and discouraged from political participation as they are unable to determine what information to trust and furthermore undermines trust in their own epistemic agency and ability to participate (Coeckelbergh, 2022). Such epistemic shame may not only discourage participation but, furthermore, encourage individuals to rely upon their own personal and emotional beliefs rather than the information they receive.

Stakeholder mistrust and miscommunication

Increased uptake and use of Generative AI programs (e.g., chatbots) for rapid communications between stakeholders will discourage face-to-face interactions for the sake of efficiency. In doing so, however, such AI programs will form a communication buffer between stakeholders increasing misunderstanding and mistrust as people may be unable to determine how the text (e.g., email, private message) reflects the thoughts and feelings of the human themselves (Illia, Colleoni and Zyglidopoulos, 2023).

The deployment of such Generative AI programs within governance, particularly in public-facing communications practices, may sew further distrust the functioning of governments and render democratic processes more difficult.

Inappropriate deployment

Due to the high quality of AI-generated media, particularly text, the capabilities of Generative AI programs are often exaggerated and misunderstood as such programs are often labelled as "thinking machines" that understand the world in the same way that human beings do. Such a comparison is misguided in that it disregards or obscures the fact that these programs display inherent biases and inaccuracies, and lack contextual understanding of social situations.

This misunderstanding of AI programs as "thinking machines" can lead to inappropriate deployment within human social settings and practices, as well as within governance. For example, in March 2022 it was reported that for the past decade the Dutch tax administration had used an inaccurate AI program to identify fraudulent childcare claims leading to innocent families, often those on low incomes and from underrepresented communities, receiving excessive tax bills, pushing thousands toward poverty, debt and, in some extreme cases, suicide (Heikkilä, 2022).

Similarly, in October 2022 the Communications and Digital Committee of the UK Parliament House of Lords questioned an AI program dubbed "Ai-Da" as a witness as part of their inquiry into the future of the creative industries. Though Ai-Da's appearance may be seen as a political gimmick, it raises the possibility for such programs to influence policy decision-making and political issues.

Liar's dividend

Uncertainty over the capabilities of GAN technologies and deepfakes provides public figures with a so-called "liar's dividend" when video evidence presenting them negatively is published online as they can more easily dismiss or deny the authenticity of this evidence (Chesney and Citron, 2018). Such uncertainty and opportunities for denial may further damage or undermine confidence in the judicial system.

For example, in 2019 a video was published online allegedly featuring Malaysian politician Azmin Ali engaging in sexual activity with another man, which is illegal in Malaysia and could lead to criminal charges. While the video itself has been proven as authentic, Ali denied his involvement and argued that the video was a realistic deepfake created to discredit him (Adjer *et al.*, 2019).

Perpetuating bias, inaccuracy and prejudice

As mentioned above with regards to automated text generation, AI programs can be trained on curated datasets in order to replicate particular language patterns that demonstrate bias. Even when using such programs for apparently neutral or positive applications, however, large language models will display inherent bias as their training datasets are so vast as to include discriminatory or non-inclusive language (Bender *et al.*, 2021).

For example, an AI-generated animated parody of the sitcom *Seinfeld* called *Nothing, Forever* began broadcasting 24/7 on the streaming platform Twitch with dialogue automatically generated using OpenAI's GPT-3 language model. However, the stream was suspended after characters began making transphobic statements that were unintended by the creators (Oladipo, 2023).

Though unintentional, the continued and increased use of inherently bias AI programs in political and policy communications may perpetuate bias, victimizing members of underrepresented communities and discouraging political participation.

Discouraging uptake of positive technologies

Increasing numbers of high-profile instances of misuse and abuse of GANs technologies may contribute to public mistrust of AI technology more generally, as well as organizations that use AI technologies (CDEI, 2022). Such mistrust may result in a reduced uptake and adoption of AI technologies, even those with overtly positive impacts for democracy. Please see Appendix A for a spreadsheet tracking high-profile examples of deepfake content, both negative and positive, and their notable characteristics. This spreadsheet will be continually updated throughout the SOLARIS project as new cases emerge.

3 Generative AI and Threats at the International Level

3.1 Generative AI and International Relations

Artificial intelligence is being developed during a period when the world is going through major geopolitical changes. The war in Ukraine and the sanctions against Russia have precipitated a process of de-coupling, where the world is being divided in different and mostly opposing groups of countries. These groups have different point of views about how the world should function, what the power relations in the international relations should look like and how the economic system should function. This de-coupling of the two greatest economic and political powers of the world increases the possibility of conflicts and decreases the possibility of a quick solution to the conflicts.

As they have started to have an impact on domestic politics, Generative AI have slowly started to have an impact in international relations, relations between states and developments in areas of conflict. Even though currently, Generative AI are still of a low quality to have a disruptive impact on the relations between states or non-state actors and states, nevertheless, artificial intelligence in general has already started to create competition between major world powers resulting in a new technological arms race. GANs and AI have started also to create internal and external problems being utilized in electoral campaigns, damaging the reputation of the opponent, or exacerbating tensions where they already exist. Even though the quality is still low and the detection relatively easy, Generative AI have the potential to create havoc in international relations, especially with the coming of a multipolar world. Unipolar or duo-polar systems have shown to be more stable in international relations and more able to prevent major wars, while multipolar systems in the past have ended up with major wars (Kennedy, 1987; Mearsheimer, 2014).

Generative AI and the new technological arms race

While the Cold War saw an arms race in conventional and nuclear arms, the 21st century is starting to see a technological arms race concerning artificial intelligence. The United States and China are going in different directions concerning artificial intelligence. This goes along with a de-coupling process between the United States and China (Felbermayr, Mahlkow and Sandkamp, 2023), which extends to de-coupling on strategic technologies (Hwang and Weinstein, 2022). China's attempts to rival the US hegemony in the technological sphere and the overall US-China rivalry has the potential to divide the world into antagonistic blocks vying for influence in the AI sphere and utilizing AI capabilities against each other.

A coming multipolar world, combined with the AI technological race between the major world powers and de-coupling process underway, will make it difficult to resolve major problems in the international scene. In this view, GANs and AI could be utilized to further divisions between the major world powers. The world today looks more and more like Europe before the First World War with several major world powers vying for hegemony. This multipolar system in the past have shown to be unstable without a major or two major states able to impose their will on the others and preventing major conflicts. In the world today, the United States, the European Union, China, Russia, Turkey, India, are trying to go their different ways despite the level of cooperation and coordination between them. Each of these powers has its own designs about the international system, which in many cases are incompatible with each other, increasing the potential for conflict. AI has the potential to expand the division already happening in the international relations, to embolden the capacities of the major world powers to follow their designs of the international system and consequently increase the likelihood of conflict.

The internal nature of the major world powers will play a role in the use of AI and GANs in the relations between them. Democratic countries will have little incentives to use AI against each other, while the current multipolar international system that is being formed involves authoritarian states like China and Russia, which will tend to use AI to further their designs of the international system, which are contrary to most of the democratic and liberal countries.

Generative AI and relations between antagonist states

1. Generative AI can be used more effectively to disrupt and create tensions, in cases where two or more states have different foreign orientations, or have a conflictual history between them or minimal political relations. Cases like this could be between Russia and Ukraine, Russia and NATO, where NATO is a non-state actor comprised of states which have different foreign orientations with Russia; Russia and Georgia, Armenia and Azerbaijan, Serbia and Kosovo. They can also be successfully utilized in cases where two or more states have territorial pretensions against each other. Cases like this could happen between China and India, India and Pakistan, China and Taiwan. In these cases, a momentary misunderstanding arising from Generative AI generated images and videos can create the ground for conflict.

2. Generative AI can be utilized in cases where two or more states have had a recent history of conflict between them, followed by normal but formal relations between them. Generative AI can create conditions for renewed conflicts. Cases like this are between Serbia and Croatia, Serbia and Kosovo, Armenia and Azerbaijan, etc.

3. Generative AI can be less successful in cases between states with consolidated internal democracy and a history of peaceful relations between them, as it would confirm the democratic peace theory (Kant, 1795). In this view, it is very unlikely that Generative AI, even in a high quality, can create problems between Denmark and Germany or USA and Canada or France and Belgium. States with democratic systems and constant communication between them have the capacity to avoid problems created by Generative AI

Generative AI and religious and ethnic divisions

1. Many states have internal divisions, based on religious, ethnic, and social components, resulting in civil wars, frozen conflicts or minimal relations between different religious and ethnic groups. Generative AI can be utilized in these cases to flare up the tensions. They can be utilized even by outside actors to create tensions, chaos and ultimately civil wars in other states. Fragile and weak states with a history of internal conflicts between different ethnic or religious communities combined with a history of lack of credibility between the groups are vulnerable to misunderstanding arising from Generative AI generated images and videos. Such AI generated conflicts have further eroded the stability of the state. In 2021 there were 233 such conflicts in the world (Statistica, 2021); cases like Bosnia and Herzegovina, North Macedonia, Lebanon, Nigeria, Ethiopia, etc.

2. Generative AI can be used even in states where these internal divisions exist, but which don't have a conflictual history between them, if they are used by certain ''actors'' inside or outside the state. Generative AI generated images or videos can be used to create tensions between such communities, even though without resulting in conflict between them, but creating distrust between communities. Such case could be Albania. The state has three religions which have a history of harmoniously living together, but a radicalized part of one of these religions can spark conflicts between them and destroy the state unity even in the absence of past conflicts. Real videos touching upon religious issues, from time to time have sparked heated debates pro and against.

Generative AI and the changing nature of war

1. The war in Ukraine has showed the newfound relevance of disinformation, misinformation and fake news in the event of war, creating confusion and disorientation in the enemy but also in the population in general and the audience around the world. The war in Ukraine until now has been the most *disinformed* and *misinformed* war in history.

2. Generated AI satellite images haven't yet had any effect on the current wars, but AI-generated fake satellite images that show a concentration of troops in the border, could force the threatened state to respond, even though the images are not real. Thus, generated deepfake satellite images or videos can create the impression that in an ongoing war the enemy forces are moving in a direction not anticipated, which will force the other party to increase or send troops in that direction, while the whole scenario is not happening at all. Open source information has widely used in documenting the Ukraine war. As knowledge and planning of the war relies more and more on OSINT, Generative AI generated fake satellite images could create havoc. This creates a dichotomy between the ''truth'' the government offers based on its satellite images and data and the ''truth'' the public looks at based on deepfake satellite images on open sources.

3. Another case could be generating AI images with the training of troops to cross an important bridge with vital military objectives, but which doesn't exist. Generative AI have been useful to spot objects and images from fake ones, but they can be used to create fake satellite images that AI can't recognize as fake.

4. But Generative AI in the Ukraine war have been used also to increase the tactical and intelligence capabilities of one part. Neural networks are used to combine ground-level photos, drone video footage with satellite imagery to enhance intelligence with the aim of increasing tactical advantage. Companies like Palantir have used AI software to analyse how the war is unfolding, troop movements and conduct battlefield damage assessments.

5. Generative AI are making war more unpredictable and easier to start in cases where there is already a tense situation between two or more parties, but also easier to conduct through the use of AI generated data, but ultimately unable to be won from one or other parties using the same technologies.

Generative AI and perception and misperception in international politics

Generative AI can be used not only against another state or another international actor, but they can be used also to create a false image and perception of a state in the eyes of another state. A state plans its actions against another state based on the image and perception that it has of that state (Jervis, 2017). This is, generated AI can be used to create a misleading image and perception of a state, usually to magnify its capabilities, in order to mislead the actions of another state. This situation can happen between two or more states with different foreign policy objectives, with a recent history of conflict or misunderstandings. In this view, especially small or less capable states can use Generative AI to create a different image of itself in the eyes of another state and therefore reduce the possibility of undertaking of hostilities from the other state.

4 Deepfakes and Fake News

In order to better understand the deep fake phenomenon, it is useful to clarify some concepts such as "fake", on the basis of the recent debate about fake news and post-truth.

4.1 Some basic definitions

Post-Truth

Oxford Dictionary definition says: "[an adjective] relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief. …the prefix in post-truth has a meaning more like 'belonging to a time in which the specified concept has become unimportant or irrelevant'".
There are two main aspects in this definition (and in similar ones):
- In public debate emotions are explicitly considered more important than before, at the expense of rational argumentation
- Not only emotions (in semiotics, *pathemic discourse*) are gaining ground in establishing what is true, but truth in general is much less important than before; authors observe a general and growing disinterest towards truth. In a public debate is more important to make our values or preferences prevail than to establish truth (see also Frankfurt, 2005 for the concept of "bullshit").

Misinformation and Disinformation

An important difference in the debate about fake news is that between misinformation and disinformation.

[1] Misinformation: "the inadvertent sharing of false information" (for instance, a journalistic error)

[2] Disinformation: "the deliberate creation and sharing of information known to be false"(Wardle, 2017).

Here it is clear that the parameter at the basis of the distinction is intentionality. Both misinformation and disinformation deal with the spreading of fake news, but in the former this happens in good faith, while in the latter there is an intention to deceive.

4.2 Classifying fake news

Tandoc *et al.* (2018) propose a model to classify fake news. They identify two dimensions, intended as continuums (items are distributed from high to low). Crossing the dimensions, we obtain four main categories, with fuzzy borders.
The dimensions are:
- *Facticity*, "refers to the degree to which fake news relies on facts" (satire is usually more corresponding to – even if distorting – reality than parody).
- *Author's immediate intention*, "refers to the degree to which the creator of fake news intends to mislead" (here we have, again, the parameter of intentionality: propaganda tends to give a biased or misleading interpretation of facts, while fabrication invent facts).

From these premises, Tandoc *et al.* derive the following taxonomic model:

		Author's immediate intention	
		High	**Low**
Level of facticity	**High**	Native advertising Propaganda Manipulation	News satire
	Low	Fabrication	News parody

Figure 3

The Jaster & Lanius proposal

Jaster & Lanius (2018) propose a slightly different model. They cross two dimensions, as Tandoc *et al.* Nevertheless, there is a first difference: their dimensions are not intended as continuums, but as binary oppositions (following a yes/no logic which creates well delimited categories). The second difference is in the definition of the two dimensions.
"…fake news is characterized by two shortcomings: it lacks truth and truthfulness. More specifically, fake news is either false or misleading (lack of truth) and it is propagated with either the intention to deceive or an utter disregard for the truth (lack of truthfulness)".
Thus, their two dimensions are:
- *Lack of truth*: is similar to Tandoc's level of facticity, because it differentiate between "true, but misleading statements" and "false statements".
- *Lack of truthfulness*: it might seem similar to Tandoc's "author's immediate intention", but it is slightly different. In Tandoc's model there can be a low or high intention to deceive; here the difference is between cases in which there is intention to deceive (which implies a concept of truth, even if negated), and cases in which truth is simply irrelevant (as in Frankfurt's concept of "bullshit").

		Lack of truth	
		False statement	**True, but misleading statements**
Lack of truthfulness	**Intention to deceive**	Lies (Pizza gate conspiracy)	Breitbart's report on young foreign men burning down a church in Germany
	Bullshit (no regard for truth)	Trump's claim to have the world's greatest memory	Trump's implication about a terror attack in Sweden

Figure 4

Jaster & Lanius examples are not very convincing. In addition, we have four categories but no terms to label them (an operation called *lexicalization* in semiotics). We could try to fix this revising their model (Fig. 5):

		Lack of truth	
		False statement	**True, but misleading statements**
Lack of truthfulness	**Intention to deceive**	Lies	Misleading comments
	Bullshit (no regard for truth)	Boasting	Words in freedom

Figure 5

Wardle's model

One of the most used model to classify fake news is the one developed by Claire Wardle (2016, 2017). In the 2017 version, Wardle recognizes seven types of fake news, which we may order from the lowest to the highest degree of fakeness:

- Satire or parody: no intention to cause harm, but has potential to fool
- False connection: when headlines, visuals or captions don't support the content
- Misleading content: misleading use of information to frame an issue or individual
- False context: when genuine content is shared with false contextual information
- Imposter content: when genuine sources are impersonated
- Manipulated content: when genuine information or imagery is manipulated to deceive
- Fabricated content: new content is 100% false, designed to deceive and do harm

Wardle's classification had the merit to show how "fake news" is a too generic umbrella term, which can be used to indicate too different phenomena. Moreover, it is very easy and useful in operative context. Nevertheless, it has some theoretical problems. Some types are problematic (imposter content is defined on the basis of the authenticity of the source; but can an imposter content be an accurate account of facts?). Above all, this is a classification founded on more criteria, and not always consistently applied (such as in Borges's Chinese encyclopedia of animals).

We can count at least these criteria:
- Intention to deceive and harm: isolating satire or parody (which are not conceived to deceive or harm) from all the other types.
- Source: the only case in which it is relevant is the imposter content (see above); by the way, from a semiotic point of view this concern more the enunciative level than the enunciate one.
- Factuality: i.e. the eventual presence of a bulk of accurate and true facts; in this case we have on one side misleading content, false connection, false context; on the other side manipulated content (where the genuine content is at this point compromised and it is difficult to isolate it from the manipulated and counterfeit one) and fabricated content. Satire/parody should belong more to this latter kind; while – as we have seen – it is not specified if factuality is relevant for the definition of imposter content.
- Degree of manipulation: the difference between manipulated content and fabricated content seems to be that the in the former there is still something related to actual facts, while in the latter everything has been generated. This criterion can be overlapped and replace that of *factuality*, now intended as a continuum from "factual but misleading content" to "manipulated content" to "fabricated content (Note: this criterion is certainly relevant for a definition of deepfakes; but are they manipulated content or rather fabricated content?).
- Various semiotic aspects: there is a bunch of three types (false connection, misleading content, false context) sharing the problem of a connection between some true facts at the core of the news and misleading elements. Now, the difference among these types should derive from differences among

these misleading elements. From a semiotic point of view we could say that false connection deals with the relations between the news content and paratextual elements (headlines, captions, etc.).
But it is hard to clarify which is the real difference between misleading content and false context, unless it is an even more confusing overlapping between the description of a mechanism (false context) and a specific intent (framing someone or an issue).
We can try to resume Wardle's model and the underlying dimensions in the following table. The "x" indicate if a dimension is relevant in defining a type and/or in distinguishing it from another one; the notes in italics indicate the value that the type takes on that dimension.

	Satire or parody	False connection	Misleading content	False context	Imposter content	Manipulated content	Fabricated content
Intention to deceive	X *No*	X *Yes*	X *Yes*	X *Yes*	X *Yes*	X *Yes*	X *Yes*
Source					X *Not authentic*		
Factuality / Degree of manipulation	?	X *More factual*	X *More factual*	X *More factual*		X *Less factual*	X *Less factual*
Text / Paratext (semiotic aspects)		X *Paratext*	X *Text (but how?)*	X *Text (but how?)*			

The table shows that Wardle's model, despite its undeniable usefulness in articulation the too generic concept of "fake news", has some degree of internal incoherence and it lacks in clarity in some definitions. Deepfakes increase the complexity of this scenario and foster the need of new models, more capable of distinguishing among different kinds of phenomena. Wardle's model, and other classifications, can still be used in research and public debate, but we could also benefit from new models, built from other points of view.

4.3 From fake news to deep fake

Intentionality is a recurring parameter in recognizing and classifying fake news. Nevertheless, even if plausible from a theoretical point of view, it is hard to use it operatively. It is very difficult, indeed, to ascertain the real intention of the author (and sometimes, his/her identity too).

In case of manipulated or fabricated content, it is easier to recognize textual traces of his/her intention to make evident the forged nature of the text itself. So the fabricating intervention of the message sender can be explicit or implicit (if the traces are hidden). On the other side we should observe the reaction of the receiver (directly – for instance through an interview – or indirectly – through report about general reception of a news), in order to know if a manipulated/fabricated content has been recognized. The work to be done on the receiver can be harder than that on the textual traces, but it is easier than ascertaining the author's intention.

From these two binary dimensions we derive a model, describing four situations we could have (Fig. 6):

		Sender	
		Explicit fabrication	**Implicit fabrication**
Receiver	**Recognition**	Contract	Unmasking
	Lack of recognition	Accident	Deception

Figure 6 - Four different kinds of deep fake exposition.

This scheme helps us to identify the main actions that should be put in place in order to avoid a harmful effect of fabricated contents such as deepfakes:

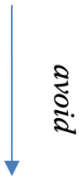
		Sender	
		Explicit fabrication	**Implicit fabrication**
Receiver	**Recognition**	Contract 	Unmasking 
	Lack of recognition	Accident	Deception

Figure 7 - Action to be done for a safer experience with deepfakes.

This means that there are cases in which explicit deepfakes are misinterpreted as real (we should investigate why this happens): we should avoid this and take appropriate precautions. In case of intentional misleading deepfakes, on the contrary, we should foster the recognition of them, going from the deception to the unmasking.

4.4 A perspective on debunking and fake news

In recent years, the creation of software that leads to the proliferation of visual disinformation - not simply verbal like most of fake news - has led to the emergence of certain professional figures such as the debunker and the fact-checker. Both these figures did not originate with fake news, but well before. In the first case, the debunker is associated with a practice of ancient origins, debunking, i.e. the person who refutes false and anti-scientific news and claims.

The modern connotation of the term debunker is attributed to journalist William Wood, who outlined its main characteristics in his 1923 book *Bunk* (Woodward, 1923). As for the figure of the fact-checker, on the other hand, it finds its origins in the editorial world from the 1920s, also in the United States. Time (1923) was the first American news weekly to set up a department dedicated to checking all the facts and information contained in the news before it went to press (Maistrello, 2013). These two roles thus seem to be two sides of the same coin. They are often associated terms, yet they are not synonymous.

The practice of debunking assumes upstream the falsity of the news, whereas fact-checking works on the news itself from a neutral point of view, verifying and confirming the facts reported in the news. It could therefore be established that since fact-checking acts to detect errors, debunking becomes a functional process. In any case, the increased virality of fake news has proportionally expanded the work of these two disinformation actors, who have had to optimize their research tools in a very short

time. For instance, *The international Fact-checking Network* was funded, which aims to develop the best practices of debunking and share them with the entire community[4].
Paolo Attivissimo is an IT journalist[5], a scholar of disinformation in the media and an established Italian debunker. During a webinar at the Zanichelli Training School[6] Attivissimo explained some of the rules of investigation that any user should know.
- *Sagan*'s *law*: extraordinary claims require extraordinary evidence.
- *Beelzebub principle*: the source of a news is more reliable if a disadvantage arises from that news for the source itself (e.g. an anticlerical news published on a Catholic newspaper).
- *Lateral Information Rule*: the most neutral information is nested in documents that talk about something else.
- *M* Mountain Principle* (not a method of investigation but an intelligent observation): making up a fake news takes very little effort, while doing a serious investigation takes a lot of effort.
Moreover, Attivissimo illustrated some of the investigation tools in the viral world, such as the advanced search proposed by Google, as well as News.google, Scholar.google or Books.google. Speaking, however, of deepfakes, the most useful open-access tool at the moment is Tineye.com[7], through which is provided a list of all sites containing a specific image or its variants (e.g. a mirror image).

Actors of visual disinformation: deepfake, debunker and fact-checking

By a search on Scopus[8] (last update 08 May 2023), with the keyword 'deepfake', the results by 2022 and 2023 turn out to be 627. There are 496 publications by 2022 and 131 by the first five months of 2023, underlining the increased scientific interest in the topic. Furthermore, we carried out other research by associating other words with the main keyword. The association with "debunking" shows only 1 article published in 2021. The one with "fact-checker" only 1 result in 2023. The combination with "risk" provides 37 results, all by 2022/23; 'impact' provides 63 results (2022/23); 'negative' provides 30 results (2022/23); finally, 'positive', provides 23 results (2022/2023).
Debunking and fact-checking have not yet been explored in depth. Which is even more unusual since these practices are now closely linked both to the digital sphere and to the semiotics approach– understanding semiotics as the discipline that studies everything that can be used to lie (Eco, 1976) and the semiotician as a scholar investigating the procedures of making something seem true (Floch, 1986).
The search yielded only two results:
a) 2021 *DDS: Deepfake Detection System through Collective Intelligence and Deep-Learning Model in Blockchain Environment* (Choi and Kim, 2023);
b) 2023, *Cutting through the Hype: Understanding the Implications of Deepfakes for the Fact-Checking Actor (Weikmann and Lecheler, 2023)* an experiment conducted by Teresa Weikmann and Sophie Lecheler in 2022 and published in 2023.
The first study (a) proposes to test whether a simple priming of information regarding deepfakes can actually improve users' ability to recognize their application in the media environment, demonstrating that a simple, yet reasoned, digital literacy intervention can lead to counteracting the deceptiveness of deepfake applications. This study was conducted in 2021, when the video images generated by

[4] For more information: https://www.poynter.org/ifcn/ - last accessed 17/07/2023

[5] Personal blog: The Disinformatic, https://attivissimo.blogspot.com/ - last accessed 08/05/2023

[6] https://www.youtube.com/watch?v=BjugRlwzSUs&t=3127s - last accessed 17/07/2023

[7] TinEye Reverse Image Search - https://tineye.com/ - last access date 02/05/2023

[8] Scopus, https://www.scopus.com/search/form.uri?display=basic#basic - last access date 08/05/2023

software producing deepfakes were not yet of high quality, making them unmaskable as fakes without too much effort, which is increasingly difficult at the present day.
The second study (b) presents an experiment based on the analysis of the impact of deepfakes on fact-checkers through ANT theory (Law, 1992; Somerville, 1997; Latour, 2007), which allows to contextualize deepfakes as a new actor entering an established and growing network of actors specializing in the detection of misinformation.
The ANT, moreover, identifies technology as an integral part of society and models its impact by also considering mutual relationships. In other words, it states that, on the one hand, deepfakes may indeed alter the way fact-checkers work, but, on the other hand, that deepfakes may also lose their threatening potential as detection techniques improve.
This study constructs a theoretical framework - inspired by the ANT - on the impact of deepfakes on fact-checkers and explores this network through the use of interviews. In order to better understand the results, it is important to emphasize that the scholars used the term misinformation, as the construction of deepfakes requires conscious action - unlike, for example, fake news. The experiment took place by interviewing 15 fact-checking experts specialized in different aspects of visual disinformation, who were required to answer mainly two questions:
- "What dangers do fact-checkers identify in connection with deepfakes for journalism and news audiences?"
- "To what extent do fact-checkers adapt their work routines to deepfakes in terms of detection techniques, and countermeasures?"
Such results suggest that today deepfakes are seen as a problem of the future, not imminent, and only alter practices to a small extent. Instead, other forms of visual disinformation such as decontextualised videos constitute a larger challenge to fact-checkers' day-to-day work, demanding the development of new manual detection processes and more specialised training for journalists. Based on these findings, this article critically discusses current journalistic discourse about deepfakes as a 'dangerous' technology and provides nuanced suggestions for future research on the phenomenon (Weikmann and Lecheler, 2023).
As anticipated, these are interviews conducted during the past year. In just a few months, these technologies have evolved exponentially, especially working on both the audio and video quality of the deepfakes, making their exposure truly deceptive.
At this point, however, it is important to bear in mind one last experiment, also conducted in 2022. Through an online survey, it was shown that informing participants about deepfakes did not improve their detection ability, but led them to believe that the videos were fake, even when they were real (Ternovski, Kalla and Aronow, 2022).
Finally, two sites designed to educate on the identification of deepfakes were tracked down within the search:
- https://www.spotdeepfakes.org/en-US (last accessed 17/07/2023).
- https://detectfakes.media.mit.edu/ (last accessed 17/07/2023).

5 A Cross-Legal Approach to Generative AI: EU Level

The issue of Generative Adversarial Networks (GANs), raises the regulatory need to integrate different legal approaches. Generative AI technology triggers more than one single specific legal problem and raises the concern of a diversity of regulatory areas. This is increasingly important considering that there is an excessive proliferation of generated AI content and synthetic media, often difficult to categorize or conceptualize. Even when the way we transmit information is classified, and fakes are categorized the main problem for the regulator stands. The current EU regulatory regime it is not sufficient to address the potential harms of a technological instrument that is used in a very democratic manner through a range of possibilities going from a modest sophisticated mechanism to a wide range of more inaccessible and mature AI technology. As part of its digital strategy, the EU is willing to regulate artificial intelligence (AI) to ensure better conditions for the development and use of this innovative technology in a safe manner The starting point in this field must be the European Commission proposal for a regulation for Artificial Intelligence. This regulation was originally proposed by the European Commission in April 2021. Beyond this proposal, it is also important to consider the Coordinated Plan on Artificial Intelligence of 2021. After the Commission proposal, the Council has adopted a common position on the AI proposal for a regulation on December 6, 2022. On June 14, 2023 the European Parliament has adopted its negotiating position to make substantial changes to the original European Commission Proposal.

Thus, although it is still soon to know what the final legal substance of the so-called AI regulation Act will be, the system proposed by the EU Commission is based on a categorization of risks. The EU Commission proposes to regulate these risks with a different level of intensity depending on the gravity of the risk. The categories are the following ones: 1. Unacceptable risks; 2. High risks; 3. Limited risks; 4. Minimal risks. The first type of risks would be banned, and the later type of risks would not require further regulation according to the proposal. The professed aim of the proposal is two-fold: to promote excellence and trust. The excellence aspects of the proposal relate to the achievement of internal market in the field of AI. Instead, the trust aspects of the proposal aim at counteracting the widespread feeling of distrust that people have vis-à-vis AI in the EU.

The intentions of the European Commission are however clear: it is difficult to achieve an internal market in the field of trust is no-one trusts AI devices. However, here the means do not justify the end. If we want good regulations in the field of AI, we must use the analytical tools that we have at our disposal carefully. It is therefore very possible to speak of reliability and efficiency in AI devices, while speaking about trust regarding the “products” that come from AI devices, instead of speaking of trust in AI in *genere*. This is the approach that will be followed in this project.

The effective problem at issue is the effective trust in the outcomes of artificial intelligence not the relevance of each of these specific applications but their potential to create a whole ecosystem of false information in a variety of forms (video, audio, photo, etc). In legal terms, this means regulating a myriad of risks from AI systems. The enormous potential to have direct and damaging impact on the rule of law, as well as society, democracy and social and professional relationships is still undetermined and not properly understood. We could speak of a handful of positive actions of GANs, but the European legislator and the policy makers are concerned with the need to equip our current regulatory systems to tackle the problems that may arise touching upon many different legal areas where potential risks may arise. The main trend has been looking at privacy and data protection law too closely (see further, Van Der Sloot and Wagensveld, 2022), but according to the latest efforts of the EU Parliament to integrate in AI legislation the need to regulate the risks deployed by the use of

generative artificial intelligence, that was not initially considered, there is enough consensus now to understand that other type of regulation and governance is needed to ensure safe management and appropriate consent for any unexpected consequences, and of course the AI Act has submitted AI transparency requirements at first instance. The other important issue at regulatory level regards the notion of trust presenting the risk of a buzzword within the AI system field without providing proper answers (Estella, 2023). Still, Generative AI could potentially affect negatively safety or fundamental rights concerned areas that have already been qualified by the EU legislator as high risk.

5.1 Possible regulatory paths

Fields of Law and Eventual regulatory Paths

Considering that we are opting for a broad legal definition it is difficult to set legal boundaries yet. In view of this, the Commission found that the notion of AI system should be more clearly defined to allocate legal responsibilities under the evolving regulatory framework. Still, it is important to understand that the dilemma that regulators are confronted with in general terms and in particular, is the question of legal certainty in this field without stifling the development of what a very promising market is. GANs and Generative AI in general touch upon a full spectrum of legal issues that are not clearly bounded at first sight. In today's complexities the use of generative AI triggers problems that need legal solutions from Public and Private Law indistinctively. A cross-legal analyses integrating different legal areas becomes particularly relevant within the context of regulatory AI systems in which the effects, not just the instrument itself could become a detriment: 1. To the individual; 2. to democracy and society as whole. Thus, a stronger public policy approach considering tailor-made mechanisms with preventive action has already been devised in view of the risks AI systems pose. Considering that the EU proposal has considered Generative AI as "limited" risk, it is important to explore whether this categorization of Generative AI as a limited risk is appropriate. In any case, we will propose to take a different regulatory path at least in the field of Generative AI. The first stance in this regard would be to propose the adoption of, at least, a directive that would be linked to the future regulation specifically on Generative AI.

Thus, we will not set boundaries here to discuss the issues for which a cross-legal analysis is needed, this is currently an open-ended regulatory process in which the first attempt has been the classification of potential risks whilst introducing measures to support innovation. We must explore is whether this categorization of Generative AI as a limited risk is appropriate or not and, if we understood that it is not, what the appropriate category for Generative AI would be. We are increasingly aware of how much a stronger regulatory approach is needed to establish a balance between beneficial uses of data and the protection of privacy, non-discrimination and other legally protected values, such as, equality, human dignity, trading possibilities, manipulations of the public, deceiving, attacks on fundamental rights, violations of intellectual property rights, transparency, data accessibility, integrity, trust, etc, We do not want this list to be exhaustive yet.

Our suggestion is the adoption of at least a directive, that would be linked to the future regulation on Generative AI. This directive would: 1. Start from the analysis of the ethical and moral values that should embed any regulation in the field of AI and of Generative AI; 2. These values would inform a general framework of Generative AI. This general framework would not be very detailed and would allow developments in this rapidly changing field; 3. Therefore, an important point of our approach would be to give an ample margin of discretion to the administrative agencies that would be put in place in the field of GANs; 4. In this regard, it will be explored the extent to which it would make

sense to set up an EU agency on AI; 5. This new EU agency would be at the top of the national agencies that the Directive would instruct to set up in the Member States.

A Comparative Look

A comparative look is crucial to contextualize better the EU setting and understand better how the world is reacting to this phenomena. This project will explore the regulation and in general terms the legislative frameworks in Generative AI in the USA, China, India and the UK. There is no common approach to AI regulation worldwide but the market is growing quickly. The extremes of the segment are formed, on the one hand, by the EU that seems to opt for a rule (legislative) based on top-down approach, general regulation of AI, and on the other hand, by countries like the UK and probably India, which have opted for much more flexibility in this field. The state-of-the-art is very much in flux, and the regulatory evolution will very much depend on the general approach to regulation of AI that each of these countries or economic areas will adopt. This project will be monitoring such regulatory trends and evolution.

A) United States
A very good summary of the current situation in the United States at the Federal level is given by this article, published in the New York Times in July, 2023. The AI is very much under discussion in the USA, a country where no specific regulation whatsoever has been yet adopted in this field. The article also points at the fact that the USA is behind the European Union when it comes to AI regulation. Apparently, the philosophy in the USA is that regulating AI could severely hinder the development of a very profitable market. The USA is the global champion in this market in competition with the Chinese. The risk of "going for private commitments" and self-regulation of the main economic players in this market is very much in place, despite the fact that some of these key players (Sam Altman, the Chief executive of Open AI, in particular) have asked to the Congress to be regulated. Concerning the State's level in the USA. Here progress is more important than the one that has been achieved at the Federal level. According to this Brooking Institution Policy brief, the State that has made more headway in this terrain has been California, which has adopted the California Artificial Intelligence Accountability Act in September 2022. Other States have adopted or updated more sectoral legislation in this field.

B) China
China is possibly one of the countries of the world that has taken regulation of AI (and regulation of Generative AI) more seriously. In this policy paper the Carnegies Endowment for International Peace, one can find a very good summary of what China is doing at this regard (Sheehan, 2023). The Chinese regulatory approach to AI seems to be more inductive and bottom-up than deductive and top down. With this we mean that the Chinese authorities have adopted a number of sectoral administrative regulations (i.e. in the field of GANs) to then check how they work in practice. After this sequence of particular and sectoral administrative regulations, it is now drafting a general law on AI. The different pieces of this regulatory framework are tied up together through a crucial document: the 2017 New Generation AI Development Plan (Robert, et al. 2021). The Chinese Government has a very clear understanding of a "first-mover" advantage in the regulatory field.

C) India
India is probably the country of the world that has adopted the most aggressive strategy regarding the deployment of AI. At present, there is no specific regulatory framework in the field of AI in India. The Indian Commission NITI Aayog (a governmental agency entrusted with think-tank tasks) has produced a number of reports (2020, 2022, 2022) on the matter, in

which there is a call for setting up of a regulatory framework in the field of AI. There is an important on-going discussion at the moment.

D) The UK

One of the sectors in which we can verify the direct impact of UK's exit from the EU is precisely the field of AI. The UK approach to AI regulation clearly differs from the one that has been adopted by the EU. The UK government published in March 2023 its White Paper on AI. Building trust in AI is cited amongst the main aims of the UK's regulatory approach to AI. In this White Paper, the UK government makes it clear that it opts for a principled approach to regulation of AI. This principled approach means that the government will only issue principles that will be addressed to the different UK agencies that deal directly or indirectly with AI. These principles are: safety, security and robustness; transparency and explainability; fairness; accountability and governance; contestability and redress. UK agencies will, in turn, adopt regulatory standards for each of the sectors that fall under their respective sphere of competence. Therefore, the UK's approach is non-statutory and sector-by-sector. Norms in place will be administrative norms. The White Paper says that it aims, with this approach, to give flexibility to the field of AI. Instead of top-down, regulatory (legislative) approach, as the one that has been adopted in the EU, the UK has adopted for a sectoral approach in which regulatory (administrative) standards will be adopted on the basis of the specificities of each sector and the principles that will be issued by the UK's government on governance of AI.

5.2 Remediation measures

The enforcement and remediation measures that should be advanced are not limited to the sectorial and disciplinary approach with which AI issues have been dealt with by lawyers, public and private lawyers alike. We will be focusing on Public Policy regulation (administrative) without losing sight of case-law development. IP issues will not be under our main concern.

We acknowledge that remediation measures should introduce deterrents through stronger public governance mechanisms that help to reduce democracy attacks, reputation implications, financial implications, education matters, and in general unexpected risks. Thus, assessment and mitigation measures should focus on the potential future impacts of the GANs and allow for continuous adaptation of the regulatory strategies.

The current transparency rule of the EU AI regulation has not dealt with an instrument that is technically unlimited. Transparency rules face the challenge of hardly undefined algorithm responses, and the need for advanced regulatory measures is evident. Instead, the European Parliament has included an express mention to Generative AI and has submitted this type of AI to transparency requirements, such as: 1. Disclosing that the content was generated by AI; 2. Designing the model to prevent it from generating illegal content; 3. Publishing summaries of copyrighted data used for training.

Furthermore, the need to think of other undefined risks triggered by AI systems has fostered policy action to design new governance and monitoring mechanisms, including a national supervisory authority, for corrective and prohibitive measures. Given the fact that this type of regulation at national and EU level takes time to take effect and assess technicalities that may affect several dimensions of our socio-economic and political life, the adoption of voluntary codes of conduct and AI Pacts are envisaged to mitigate the potential downsides of Generative AI. This specific issue will be addressed and revised during the project to understand better remediation and enforcement measures.

6 Generative AI and The Media

6.1 Mapping spread of deepfakes in social media and response of media: Bulgaria as case study

For the sake of efficiency, the research team resolved to zoom into a particular case of distribution of deepfakes in the media – Bulgaria. The media landscape in this EU member is a particularly interesting use case when it comes to the distribution of disinformation, and this is so for several reasons.
One of them is the society's consistently low media literacy score and its high vulnerability to disinformation campaigns, as demonstrated by the Media Literacy Index of Open Society Institute - Sofia. The 2023 edition of the Index[9] once again ranks Bulgaria as the least media literate nation in the EU and one of the most vulnerable countries to disinformation in Europe. Although the country is doing worse than anyone else in the EU, it is still closely resembling and to some extend typical example for the region – of South-eastern Europe which generally performs worse than Northern and Western countries. The authors of the Index place Bulgaria in the same cluster as Romania, Moldova, Serbia, Montenegro, that are all classified as problematic.
Another reason for Bulgaria's research value is its close links with Russia, regarded in the EU as one of the most prominent sources of fake news, manipulative contents and disinformation.
On top of language similarities (both countries use languages of Slavic origins and the Cyrillic alphabet, which considerably facilitates the understanding between the two, simplifying the distribution of content in Russian and its understanding and uptake by the population without the need of translation), there is a very strong cultural and historical tie between the two. Russia is still regarded by many as the Liberator of Bulgarians from the 5-centuries long dominance of Ottoman empire in late 19th century. For this reason, large portions of society look at Russia with gratefulness, trust and are prone to be less critical to anything that comes from Kremlin. The common socialist past of the two countries is another reason for the population that is still nostalgic to this period to be more prone to believing Russian-version of events. In this regard, Bulgaria might also be considered as similar to other former socialist countries, such as Hungary, Slovakia.
In a study by the global think-tank GLOBSEC from 2022[10], shortly after the start of Russia's invasion of Ukraine, the positive attitudes of Bulgarians towards Russia and the dominant opinion that one should maintain neutrality towards the war are clearly visible. Only in Hungary and Slovakia are there more positive attitudes towards the war, Russia and Vladimir Putin.
Finally, the lack of media pluralism is obvious and the high concentration of media outlets into few big groups distorts the viewpoints, controlling the exposure to information content of Bulgarians. This effect is further enhanced by the domination of Facebook – a social network, criticized for the augmentation of echo chambers effects and for its failure to tackle the issue of disinformation spread. We believe that all of this makes Bulgaria a favourable ground for the development of fake news and deep fake content generated by adversarial networks and used for further manipulation of public opinion. A direct consequence of this might be the hampering of Bulgaria's industrial capacity to help Ukraine with ammunition and serving Russia's interests to quickly take over Ukraine.

As for the methodology, the mapping of GANs activity in social media in Bulgaria is based on expert observation and desk research of social media and fact-checking platforms. The author, based on his professional experience in the news media sector and knowledge of the social media environment could not retrieve any GANs-generated deepfakes circulating in the Bulgarian social media environment, let alone such originating from Bulgaria**.**

[9] https://osis.bg/?p=4449 – last accessed 28/08/2023

[10] https://www.globsec.org/sites/default/files/2022-05/GLOBSEC-Trends-2022.pdf last accessed 1/09/2023

These observations were cross-checked through consultation with other Bulgarian professionals in the field of journalism and the results were the same. To complement this, the author performed desk research of the social media environment through digital tools and a search on the most authoritative fact-check platforms to date – Agence France Presse's Fact Check Proveri.afp.com[11] and factcheck.bg[12]. The last step was performed to define if there have been some impactful GANs generated contents that impacted the media environment (since a main principle of such platform is to fact-check only news-worthy content that is likely to be falsified).

The scope of the research was determined afterwards. It was mainly aimed at the social networks belonging the Meta Platforms group, which are the most popular in Bulgaria. Based on Napoleon-Cat's data, the most popular social media channel in Bulgaria is Facebook, followed by Instagram and Messenger[13]. In particular, national statistics show that Facebook is used by nearly 85% of Bulgarian citizens with internet access[14].

The huge reach of Meta in Bulgaria determines the significant interest in the creators of fake content, disinformation and some types of deep fake to use Facebook for its spread.

Similarly, the analysis of some of the most popular and distinctive Facebook groups and pages[15] known to be used to create and disseminate fake news and disinformation in Bulgaria over a two-year period shows that Bulgaria currently lacks widespread high-quality deep fake content that is almost indistinguishable from authentic content (video and photos). Based on this, it can be concluded that GANs are currently not used on social networks to create content that pretends to be authentic.

However, manipulated videos are a widespread way to create disinformation. The most popular way is through editing of excerpts from existing videos aimed at telling some falsified story by real people[16]. This usually happens by placing subtitles that are either not related to the real statement of the person from the video or are taken out of context[17].

Most often, this form of disinformation is used for the benefit of Russia: for instance, using videos of Western experts in a particular field or officials who are portrayed as recognizing a particular conspiracy theory (proliferation of COVID-19, chemical weapons, HAARP, chemtrails, interference in elections or biological experiments)[18]. Such videos are distributed mainly in Facebook groups and pages, with an obvious pro-Russian bias. They spread on other social networks, but occasionally – in the Reddit – in the Subreddit for r/Bulgaria[19]. They are not covered in the mainstream media and in the more serious and legitimate online media.

In Figure 8 (below) US journalist and writer Stephen Kinzer appears in a manipulated video spreading on Bulgarian social media. The subtitles read "Yes, it is true. We still interfere in foreign elections. The last I can think of are those in Bulgaria".

On the right-hand side of the screen, can be seen Facebook groups where the video was shared – mostly with pro-Russian or conspiracy profile.

[11] https://proveri.afp.com/list - last accessed 17/07/2023

[12] https://factcheck.bg - last accessed 17/07/2023

[13] https://napoleoncat.com/stats/social-media-users-in-bulgaria/2023/04/ - last accessed 17/07/2023

[14] https://www.nsi.bg/bg/content/2808/достъп-на-домакинствата-до-интернет - last accessed 17/07/2023

[15] https://www.facebook.com/profile.php?id=100064622146606 - last accessed 17/07/2023

[16] https://www.facebook.com/permalink.php?story_fbid=pfbid022PrJUnZvTWsj4jV6h4Wgqe1NDn5vjh7CjKask4qzmbhW3DoYEHtoc6ivqkRS6xDsl&id=100064622146606 - last accessed 17/07/2023

[17] https://www.facebook.com/watch/?v=1455768331607046 - last accessed 17/07/2023

[18] https://www.facebook.com/groups/479228473907715/permalink/705611647936062 - last accessed 17/07/2023

[19] https://www.reddit.com/r/bulgaria/comments/119uhtv/да_вярно_е_ние_сащ_все_още_се_бъркаме_в_чужди/ - last accessed 17/07/2023

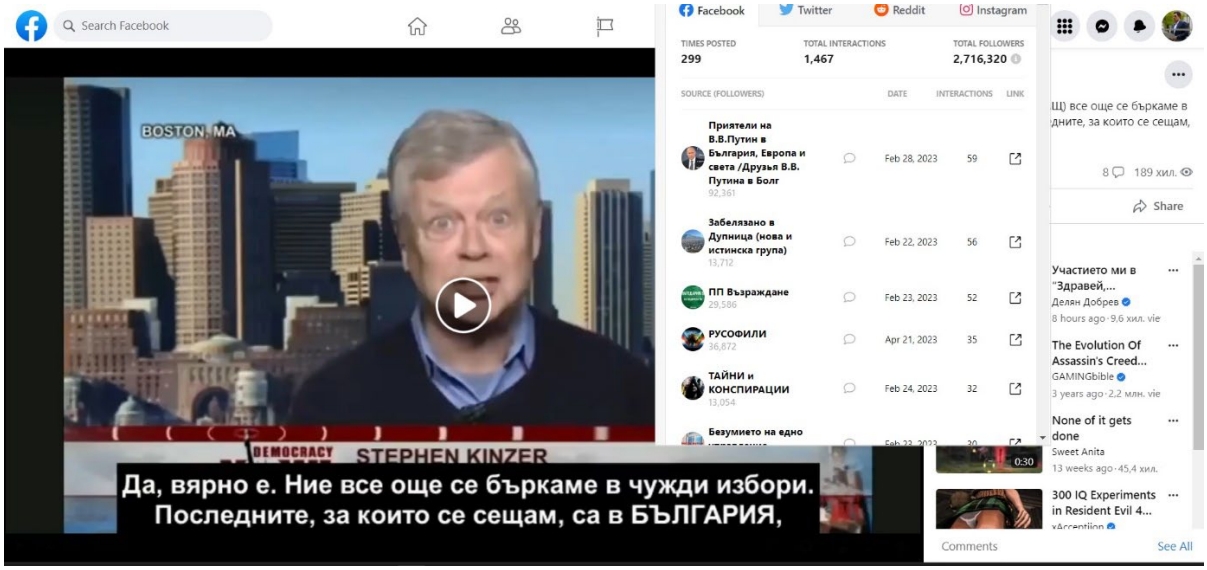


Figure 8

In the following paragraphs we will illustrate the dominant manipulative content in social networks in Bulgaria at the moment.

6.2 Examples of manipulated deepfakes and their spread to Facebook groups and pages

U.S. Interference in Bulgaria's Elections

One of the most popular manipulated videos among Bulgarian Facebook users[20] according to the Crowd Tangle platform, is claimed to be a proof of US interference in the recent elections in Bulgaria. The post began to gain popularity in February. According to the CrowdTangle tool, it has been posted 299 times in groups to date, in which there are over 2.7 million followers in total.

The video is in fact an excerpt from the American show Democracy Now! from 12 March 2018, but it is shared without specifying the publication date and suggesting that these are recent events. The original interview lasts 8 minutes, while what is shared on Facebook is only 4 minutes and 39 seconds. The excerpt is additionally edited. In the authentic version, Bulgaria is mentioned in the course of the conversation, and in the video that spreads on Facebook, the sentence with Bulgaria is cut out and placed at the front.

In the manipulated version of the interview, the context of the conversation is not clear. In the real video, the topic is the interference of the US in the electoral processes in other countries historically, and the reason is the ongoing investigation into possible Russian interference in the US elections in 2016. Further confusion comes from the fact that two people spoke in the excerpt – the interview was with writer and journalist Stephen Kinzer, but at the beginning of the conversation a recording of former CIA Director James Woolsey was broadcast on Fox News a month earlier. Asked if the U.S. had tried to interfere in foreign elections, Woolsey confirms and gives as an example "Europe in 1947, 1948 and 1949". He explained that the goal was not to bring the communists to power.

[20] https://www.facebook.com/groups/708111179732964/permalink/1377759062768169 - last accessed 17/07/2023

Conspiracies for the HAARP program

Another case in which the name of a former FBI director, Ted Gunderson, was used again. The video[21], which is of low quality and difficult to tell whether even this is Gunderson himself, talks about the American High Frequency Active Auroral Research Program (HAARP). This is an existing radio frequency research program, but the manipulated video states that it is about chemical attacks on citizens.
The entire video is created by compiling different individual videos, for which it is unclear whether they are on one topic at all. The video itself contains speeches by various American experts, activists and scientists, but they are edited chaotically. The voice-over most often instills the guilt of the United States for various problems in countries such as Iran, Venezuela and even the Soviet Union.
The video is posted on a page that is apparently designed solely to spread disinformation[22]. It was published on April 30, and within 4 days it is already shared in groups with nearly 600,000 followers.

Vaccines are poisoning us

Another example is a video[23] claiming that Pfizer's vaccine has a negative impact on human blood. In this case, a video with a real doctor whose voice is used, and he explains about some of his research. Against the backdrop of this video and voice, however, subtitles are placed that do not correspond in any way and are intended to manipulate viewers. The video is again distributed in hundreds of groups and pages.

Suspicion of deep fake with political repercussions

Observations show that deepfakes are not used to deal with political opponents in Bulgaria. The only case of suspected occurrence is via supposedly manipulated images and video from the bedroom of Bulgarian Prime Minister Boyko Borisov. In 2020, photos of unclear origin appeared in the public news space, portraying the sleeping Prime Minister, with his open locker full of euro banknotes and gold bars (Fig 9).

[21] https://fb.watch/udyeM2HRgz/- last accessed 27/08/2024

[22] https://t.me/myfirstvideospot- last accessed 27/07/2024

[23] By the time of writing the deliverable, the video was available on this link: https://www.facebook.com/groups/2676054155849327/posts/6446050385516333/ and distributed through the group of Plamen Paskov here: https://www.facebook.com/groups/2676054155849327, among other locations. It has since been taken down, unavailable anywhere on the Internet which suggests a moderation action has been taken by the Facebook platform.

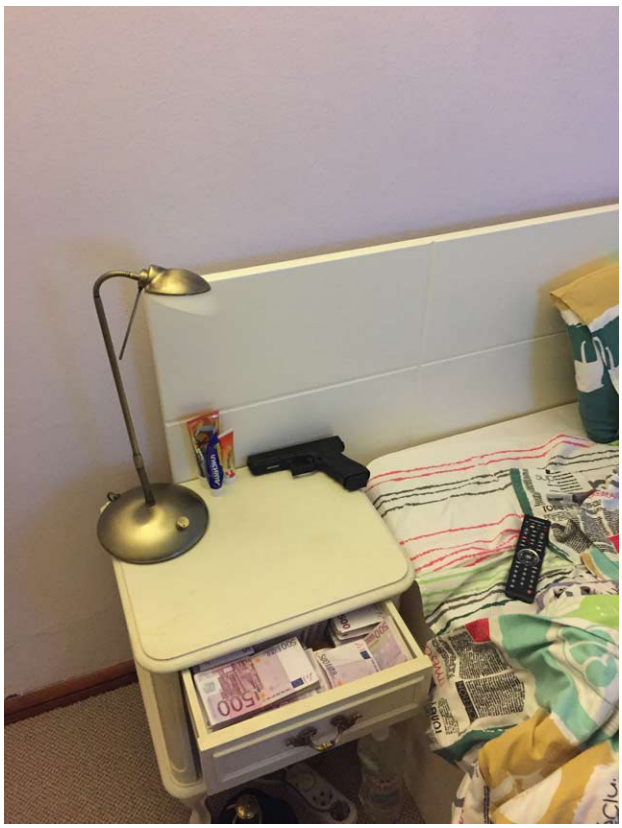


Figure 9 - An image of a locker, full of euro banknotes, allegedly from the bedroom of Prime Minister Boyko Borisov. The publication is from 2020. Source: Unknown

He himself denied that the photos were real claiming that they were fakes, but several independent examinations[24] [25] proved him wrong.
The Bulgarian prosecutor's office, which many analysts accuse of close ties to Borisov, cleared him by saying the photos were indeed manipulated[26]. There is still no finalized expertise for the origin of video and the technology used to manipulate it, if any at all.

Humoristic deep fake

Another genre of widespread deepfakes in Bulgaria are those of humoristic nature. Although they present themselves as funny pictures in which the faces of certain characters are replaced, mainly, with political figures, their goal is to discredit political opponents. For their dissemination are used mainly influencers[27] (Fig. 10 and onwards).

[24] https://e-vestnik.bg/32531/snimkite-kompromat-za-borisov-sa-praveni-2017-2019-bez-fotoshop-bez-montazhi-korigirani-sa-datite/ - last accessed 17/07/2023

[25] https://e-vestnik.bg/36066/ekspertizata-na-geshev-kakvo-izlagaha/ - last accessed 17/07/2023

[26] https://www.mediapool.bg/prokuraturata-onevini-borisov-za-snimkite-ot-spalnyata-chrez-evrodeputat-news337650.html - last accessed 17/07/2023

[27] https://www.facebook.com/TheRedIvan - last accessed 17/07/2023

Figure 10 - The face of the leader of the leading political party GERB Boyko Borissov on the body of TikTok star Chechenetsa. Source: Ivan Chervenkov on Facebook.

Figure 11 - The face of former Bulgarian Prime Minister Boyko Borisov superimposed on an image of Venezuelan President Nicolas Maduro. Source: Ivan Chervenkov on Facebook

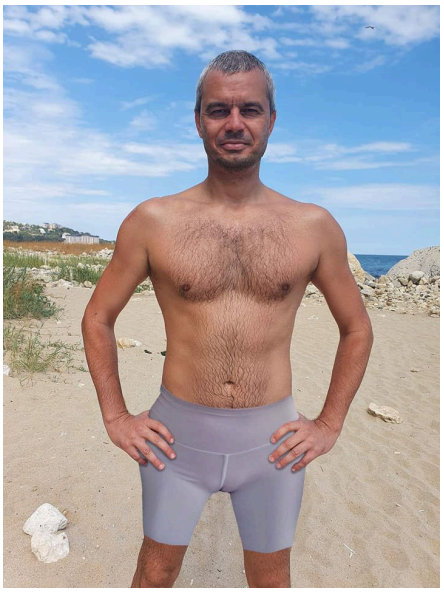


Figure 12 - Photo: Collage with the leader of the pro-Kremlin party "Vazrazhdane" Kostadin Kostadinov. Source: Ivan Chervenkov on Facebook

Figure 13 - Photo: Another collage with the leader of the party "Vazrazhdane". Source: Ivan Cher-venkov on Facebook

Figure 14 - Photo: The Bulgarian Chief Prosecutor as an astronaut. Source: Gospodari on Facebook

Figure 14 - The face of Bulgarian President Rumen Radev imposed on his opponent Boyko Borisov's body, hinting at a symbiosis between the two. Source: Ivan Chervenkov on Facebook

Figure 15 - L-R: The leader of the Bulgarian Socialists Kornelia Ninova, the leader of GERB Boyko Borisov and President Rumen Radev as part of the plot of the film *Forrest Gump*. Source: Ivan Chervenkov on Facebook.

Figure 16 - Bulgarian Justice Minister Krum Zarkov in a fictional skirmish with Dutch Prime Minister Mark Rutte over Amsterdam's refusal to allow Sofia into the Schengen area. Source: Ivan Chervenkov on Facebook.

Figure 17 - Ridiculing of Russian Foreign Minister Sergey Lavrov. Source: Ivan Chervenkov on Facebook

Figure 18 - Photo: The face of the former Minister of e-Government Bozhidar Bozhanov, an IT specialist, on the poster of the film "The Imitation Game". The reason are the accusations by the opposition that he ma-nipulates the voting machines in Bulgaria

Institutional response?

The mass and uncontrolled spread of Russian propaganda and disinformation is becoming acknowledged as a serious problem in Bulgaria, mainly because of its influence during elections, which in Bulgaria became more frequent. One of the parties in parliament submitted a bill in this direction[28], but because of the short life of parliament it was not moved forward. Electronic Government Minister Bozhidar Bozhanov contacted Facebook, trying to get the social network to tighten its control over content in Bulgaria – in particular, to analyze Bulgaria as a use case and improve based on it, but the company refused[29].
However, there was an effect of the raised topic – Facebook terminated the contract with the company, responsible for moderating Facebook in Bulgaria – Telus International – which has received public accusations of having a pro-Russian approach.
Regarding the spreading manipulations, there was only reaction to the photos and videos of former Prime Minister Boyko Borisov, because they affected him personally. The prosecutor's office is acting in his favor.

6.3 Conclusion

Bulgaria is a “battleground” for fake news, manipulation and disinformation on the Internet. For now, however, the tools used are low-quality and easily identifiable. There is no known occurrence of deepfakes that is good enough to be unrecognizable. The reasons may be several: high quality requires too many resources, which in Bulgaria at this stage are not justified. Manipulation still appears to be enough. In addition, from 2020, Facebook has introduced mechanisms to combat deepfakes, and the risk of it being detected quickly makes the creation of such videos and photos less profitable.

[28] https://offnews.bg/politika/db-predlaga-kak-da-se-ogranichat-trolovete-po-vreme-na-predizborna-789278.html - last accessed 17/07/2023

[29] https://www.economic.bg/bg/a/view/facebook-otkaza-da-izpolzva-bylgarija-za-izuchavane-na-dezinformacijata - last accessed 17/07/2023

References

Adjer, Henry et al. *The State of Deepfakes: Landscape, Threats, and Impacts*. Deeptrace (Adjer, H. *et al.* (2019) 'The State of Deepfakes: Landscape, Threats, and Impacts'.

Ascott, T. (2020) 'Microfake: How Small-scale Deepfakes Can Undermine Society', *Journal of Digital Media and Policy*, 11.2, pp. 215–222.

Attivissimo, P. (no date) *Il Disinformatico*. Available at: https://attivissimo.blogspot.com/ (Accessed: 5 August 2023).

Bender, E.M. *et al.* (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, pp. 610–623. Available at: https://doi.org/10.1145/3442188.3445922.

Biswas, S. (2023) 'ChatGPT and the Future of Medical Writing', *Radiology*, 307(2), p. e223312. Available at: https://doi.org/10.1148/radiol.223312.

Buchanan, B. *et al.* (2021) *Truth, Lies, and Automation: How Language Models Could Change Disinformation*. Center for Security and Emerging Technology. Available at: https://doi.org/10.51593/2021CA003.

Burgess, S. (2022) 'Ukraine War: Deepfake Video of Zelenskyy Telling Ukrainians to "Lay Down Arms" Debunked', *Sky News*, 17 March. Available at: https://news.sky.com/story/ukraine-war-deepfake-video-of-zelenskyy-telling-ukrainians-to-lay-down-arms-debunked-12567789.

Chesney, R. and Citron, D. (2018) 'The Coming Age of Post-Truth Geopolitics', *Foreign Affairs*, 11 December. Available at: https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war?utm_medium=promo_email&utm_source=lo_flows&utm_campaign=registered_user_welcome&utm_term=email_1&utm_content=20230404.

Choi, N. and Kim, H. (2023) 'DDS: Deepfake Detection System through Collective Intelligence and Deep-Learning Model in Blockchain Environment', *Applied Sciences*, 13(4), p. 2122. Available at: https://doi.org/10.3390/app13042122.

Coeckelbergh, M. (2022) 'Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence', *AI and Ethics* [Preprint]. Available at: https://doi.org/10.1007/s43681-022-00239-4.

Damiani, J. (2019) 'James Dean To Be Digitally Resurrected To Appear In His Fourth Film, "Finding Jack"', *Forbes*, 7 November. Available at: https://www.forbes.com/sites/jessedamiani/2019/11/07/james-dean-to-be-digitally-resurrected-to-appear-in-his-fourth-film-finding-jack/.

Doherty, B. (2023) 'Manus Island and Nauru: previously unseen testimony and AI imagery reveal "unimaginable" part of Australian history', *The Guardian*, 8 April. Available at: https://www.theguardian.com/australia-news/2023/apr/08/manus-island-and-nauru-previously-unseen-testimony-and-ai-imagery-reveal-unimaginable-part-of-australian-history.

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

Estella, A. (2023) 'Trust in Artificial Intelligence: Analysis of the European Commission Proposal for a Regulation of Artificial Intelligence', *Indiana Journal of Global Legal Studies*, Vol. 30: Iss. 1, Article 4. Available at: Available at: https://www.repository.law.indiana.edu/ijgls/vol30/iss1/4.

EUvsDiSiNFO| Research Archive (no date) *EUvsDisinfo*. Available at: https://euvsdisinfo.eu/research/ (Accessed: 28 August 2023).

facebook.com/TheRedIvan (2023). Available at: https://www.facebook.com/TheRedIvan (Accessed: 29 September 2023).

Factcheck.bg – Проверка на факти (2023) *Factcheck.bg – Проверка на факти*. Available at: https://factcheck.bg (Accessed: 17 July 2023).

Felbermayr, G., Mahlkow, H. and Sandkamp, A. (2023) 'Cutting through the value chain: the long-run effects of decoupling the East from the West', *Empirica*, 50(1), pp. 75–108. Available at: https://doi.org/10.1007/s10663-022-09561-w.

Fingas, J. (2021) 'Deepfake satellite images pose serious military and political challenges', *Engadget*, 27 April. Available at: https://www.engadget.com/deepfake-satellite-imagery-144145142.html.

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

Frankfurt, H.G. (2005) *On bullshit*. Princeton, NJ: Princeton University Press.

'GLOBSEC Trends 2022' (no date). Available at: https://www.globsec.org/sites/default/files/2022-05/GLOBSEC-Trends-2022.pdf (Accessed: 29 September 2023).

Goodfellow, I.J. *et al.* (2014) 'Generative Adversarial Networks'. arXiv. Available at: http://arxiv.org/abs/1406.2661 (Accessed: 28 September 2023).

Hajdu, D. *et al.* (no date) *GLOBSEC Trends 2022*. Centre for Democracy & Resilience. Available at: https://www.globsec.org/sites/default/files/2022-05/GLOBSEC-Trends-2022.pdf (Accessed: 9 January 2023).

Harwell, D. (2018) 'Fake-porn videos are being weaponized to harass and humiliate women: 'Everybody is a potential target', *The Washington Post*, 30 December. Available at: https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/.

Heikkilä, M. (2022) 'Dutch scandal serves as a warning for Europe over risks of using algorithms', *Politico*, 29 March. Available at: https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

Hwang, T. and Weinstein, E. (2022) 'Decoupling in Strategic Technologies - Center for Security and Emerging Technology'. Center for Security and Emerging Technology.

Il Post (2023) 'Il finto messaggio di Putin diffuso sulle radio e tv russe', 5 June. Available at: https://www.ilpost.it/2023/06/05/finto-messaggio-putin/?homepagePosition=0.

Illia, L., Colleoni, E. and Zyglidopoulos, S. (2023) 'Ethical implications of text generation in the age of artificial intelligence', *Business Ethics, the Environment & Responsibility*, 32(1), pp. 201–210. Available at: https://doi.org/10.1111/beer.12479.

Indian Commission NITI Aayog (2020) *Responsible AI*. Indian Commission NITI Aayog. Available at: https://niti.gov.in/sites/default/files/2020-07/Responsible-AI.pdf.

Indian Commission NITI Aayog (2021) *AI For All*. Indian Commission NITI Aayog. Available at: https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf.

Indian Commission NITI Aayog (2022) *Responsible AI*. Indian Commission NITI Aayog. Available at: https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf.

'International Fact-Checking Network' (no date) *Poynter*. Available at: https://www.poynter.org/ifcn/ (Accessed: 17 July 2023).

Jervis, R. (2017) *Perception and Misperception in International Politics: New Edition*. Princeton University Press. Available at: https://doi.org/10.2307/j.ctvc77bx3.

Kant, I. (1795) *Zum ewigen Frieden. Ein philosophischer Entwurf*. F. Nicolovius.

Kennedy, P.M. (1987) *The Rise and fall of the great powers: economic change and military conflict from 1500 to 2000*. New York: Random House.

Latour, B. (2007) *Reassembling the social: an introduction to Actor-Network-Theory*. 1. publ. in pbk. Oxford: Oxford Univ. Press (Clarendon lectures in management studies).

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

Maistrello, S. (2013) *Fact checking. Dal giornalismo alla rete*. Apogeo.

McGuffie, K. and Newhouse, A. (2020) 'The Radicalization Risks of GPT-3 and Advanced Neural Language Models', *Middlebury Institute of International Studies* [Preprint]. Available at: https://doi.org/10.48550/ARXIV.2009.06807.

Mearsheimer, J.J. (2014) *The tragedy of great power politics*. Updated edition. New York: W.W. Norton & Company (The Norton series in world politics).

Mirza, M. and Osindero, S. (2014) 'Conditional Generative Adversarial Nets'. arXiv. Available at: http://arxiv.org/abs/1411.1784 (Accessed: 28 September 2023).

MIT Media Lab (no date) *DeepFakes, Can You Spot Them?* Available at: https://detectfakes.media.mit.edu/ (Accessed: 17 July 2023).

Niki Leaks (2022a) *Niki Leaks - ИЗВЪНРЕДНО: В изслушването за Кромид директор на... | Facebook*. Available at: https://www.facebook.com/permalink.php?story_fbid=pfbid022PrJUnZvTWsj4jV6h4Wgqe1NDn5vjh7CjKask4qzmbhW3DoYEHtoc6ivqkRS6xDsl&id=100064622146606 (Accessed: 17 July 2023).

Niki Leaks (2022b) *Video | Facebook*. Available at: https://www.facebook.com/watch/?v=1455768331607046 (Accessed: 17 July 2023).

Niki Leaks (2023) *Niki Leaks - Публикувам втората част на документалния филм 'Франкенскайс'... | Facebook*. Available at: https://www.facebook.com/watch/?v=263745229428660 (Accessed: 17 July 2023).

Niki Leaks (no date) *Facebook Page*. Available at: https://www.facebook.com/profile.php?id=100064622146606 (Accessed: 17 July 2023).

Novak, M. (2023) 'GOP Releases First Ever AI-Created Attack Ad Against President Biden', *Forbes*, 25 April. Available at: https://www.forbes.com/sites/mattnovak/2023/04/25/gop-releases-first-ever-ai-created-attack-ad-against-president-biden/.

Obvious (2018) *Portrait d'Edmond de Belamy* [Generative adversarial network portrait painting].

Oladipo, G. (2023) 'AI-generated Seinfeld parody banned on Twitch over transphobic standup bit', *The Guardian*, 6 February. Available at: https://www.theguardian.com/us-news/2023/feb/06/nothing-forever-twitch-ban-seinfeld-parody-ai.

Romy Jaster and David Lanius (2018) 'What is Fake News?', *Versus*, (2), pp. 207–224. Available at: https://doi.org/10.14649/91352.

Satter, R. (2019) 'Experts: Spy Used AI-generated Face to Connect With Targets', *Associated Press*, 13 June. Available at: https://apnews.com/article/ap-top-news-artificial-intelligence-social-platforms-think-tanks-politics-bc2f19097a4c4fffaa00de6770b8a60d.

Scopus (no date). Available at: https://www.scopus.com/standard/marketing.uri#basic (Accessed: 5 August 2023).

Sheehan, M. (2023) 'China's AI Regulations and How They Get Made'. Carnegie Endowment for International Peace. Available at: https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

Smith, H. and Mansted, K. (2020) 'Weaponised Deepfakes: National Security and Democracy', *International Cyber Policy Centre* [Preprint].

Social Media users in Bulgaria - April 2023 (no date). Available at: https://napoleoncat.com/stats/social-media-users-in-bulgaria/2023/04/ (Accessed: 17 July 2023).

Somerville, I. (1997) 'Actor-Network Theory: A useful paradigm for theanalysis of the UK cable/on-line sociotechnical ensemble?', *AMCIS* [Preprint]. Available at: https://aisel.aisnet.org/amcis1997/37.

Spot the Deepfake (no date). Available at: https://www.spotdeepfakes.org/en-US (Accessed: 17 July 2023).

Statistica (2021) 'Number of domestic and international conflicts worldwide in 2021, by conflict intensity', *Statistica*. Available at: https://www.statista.com/statistics/278212/number-of-domestic-and-international-conflicts-worldwide-by-intensity/#statisticContainer.

Tandoc, E.C., Lim, Z.W. and Ling, R. (2018) 'Defining "Fake News": A typology of scholarly definitions', *Digital Journalism*, 6(2), pp. 137–153. Available at: https://doi.org/10.1080/21670811.2017.1360143.

Ternovski, J., Kalla, J. and Aronow, P. (2022) 'Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments', *Journal of Online Trust and Safety*, 1(2). Available at: https://doi.org/10.54501/jots.v1i2.28.

ThisPersonDoesNotExist - Random AI Generated Photos of Fake Persons (no date). Available at: https://this-person-does-not-exist.com/en (Accessed: 17 July 2023).

TinEye Reverse Image Search (no date). Available at: https://tineye.com/ (Accessed: 5 February 2023).

u/Georgi (2023) *Reddit - Да, вярно е. Ние (САЩ) все още се бъркаме в чужди избори. Последните, за които се сещам, са тези в БЪЛГАРИЯ..." Само с добра кауза и в името на...* Available at: https://www.reddit.com/r/bulgaria/comments/119uhtv/да_вярно_е_ние_сащ_все_още_се_бъркаме_в_чужди/ (Accessed: 17 July 2023).

Vaccari, C. and Chadwick, A. (2020) 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News', *Social Media + Society*, 6(1), p. 205630512090340. Available at: https://doi.org/10.1177/2056305120903408.

Van der Nagel, E. (2020) 'Verifying Images: Deepfakes, Control, and Consent', *Porn Studies*, 7.4, pp. 424–429.

Van Der Sloot, B. and Wagensveld, Y. (2022) 'Deepfakes: regulatory challenges for the synthetic society', *Computer Law & Security Review*, 46, p. 105716. Available at: https://doi.org/10.1016/j.clsr.2022.105716.

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

Wardle, C. (2017) 'Fake news. It's complicated.', *First Draft*, 16 February. Available at: https://firstdraftnews.org/articles/fake-news-complicated/.

Webinar - Paolo Attivissimo - 'Il Debunking. Come si smascherano le bufale scientifiche' (2018). Available at: https://www.youtube.com/watch?v=BjugRlwzSUs (Accessed: 17 July 2023).

Weikmann, T. and Lecheler, S. (2023) 'Cutting through the Hype: Understanding the Implications of Deepfakes for the Fact-Checking Actor-Network', *Digital Journalism*, pp. 1–18. Available at: https://doi.org/10.1080/21670811.2023.2194665.

Woodward, W. (1923) *Bunk*. Harper & Brothers.

Zheng, H. and Zhan, H. (2023) 'ChatGPT in Scientific Writing: A Cautionary Tale', *The American Journal of Medicine*, 136(8), pp. 725-726.e6. Available at: https://doi.org/10.1016/j.amjmed.2023.02.011.

Вавова, М. (2022) *Facebook отказа да използва България за изучаване на дезинформацията*, *Economic.bg*. Available at: https://www.economic.bg/bg/a/view/facebook-otkaza-da-izpolzva-bylgarija-za-izuchavane-na-dezinformacijata (Accessed: 17 July 2023).

Достъп на домакинствата до интернет (2022). Available at: https://www.nsi.bg/bg/content/2808/достъп-на-домакинствата-до-интернет(Accessed: 17 July 2023).

'Експертизата на Гешев – какво излъгаха?' (2022) *e-vestnik.bg*, 21 July. Available at: https://e-vestnik.bg/36066/ekspertizata-na-geshev-kakvo-izlagaha/ (Accessed: 17 July 2023).

ЗА ДРУЖБА МЕЖДУ БЪЛГАРИЯ И РУСИЯ | Facebook (no date). Available at: https://www.facebook.com/groups/479228473907715/permalink/705611647936062 (Accessed: 17 July 2023).

'Индекс за медийна грамотност 2023: България е най-уязвима към дезинформация сред страните в ЕС' (2023) *osis.bg*, 24 June. Available at: https://osis.bg/?p=4449 (Accessed: 28 August 2023).

Пламен Пасков | Facebook (no date). Available at: https://www.facebook.com/groups/2676054155849327/posts/6446050385516333/ (Accessed: 17 July 2023).

Прокуратурата оневини Борисов за снимките от спалнята чрез евродепутат (2022) *Mediapool.bg*. Available at: https://www.mediapool.bg/prokuraturata-onevini-borisov-za-snimkite-ot-spalnyata-chrez-evrodeputat-news337650.html (Accessed: 17 July 2023).

Райчева, Р. (2022) *ДБ предлага как да се ограничат '"троловете"' по време на предизборна кампания - Политика*, *offnews.bg*. Available at: //offnews.bg/politika/db-predlaga-kak-da-se-ogranichat-trolovete-po-vreme-na-predizborna-789278.html (Accessed: 17 July 2023).

'Снимките компромат за Борисов са правени 2017-2019, без „Фотошоп", без монтажи, коригирани са датите' (2020) *e-vestnik.bg*, 21 June. Available at: https://e-vestnik.bg/32531/snimkite-kompromat-za-borisov-sa-praveni-2017-2019-bez-fotoshop-bez-montazhi-korigirani-sa-datite/ (Accessed: 17 July 2023).

Събуди се, РОБЕ!!! | Facebook (no date). Available at: https://www.facebook.com/groups/708111179732964/permalink/1377759062768169 (Accessed: 17 July 2023).

Хиляди потребители споделят измислено интервю на Мартин Карбовски с Владимир Путин (2023) *Провери*. Available at: https://proveri.afp.com/list (Accessed: 17 July 2023).

Appendices

Appendix A: Deepfake case studies spreadsheet

Categories
Entertainment: deepfake content used within an artistic production intended for public reception (e.g., film, TV)
Social: deepfake content circulated on social media platforms and across online networks (e.g., memes, videos)
Educational: deepfake content created and displayed for the purposes of education (e.g., interactive exhibitions)
Journalism: deepfake content used as a means of communicating news information
Political: deepfake content designed to promote a particular political position on an issue (e.g., disinformation)
Promotional: deepfake content used for advertising products and services
Pornographic: deepfake content used for artificial pornography
Safeguard: deepfake alteration of a witness face as a means of public/legal protecting of personal identity

Characteristics
Authorization: has the target provided consent for the deepfake?
Disclosure: has the artificiality of the deepfake been disclosed to the viewer?
Function: how and for what purpose has the deepfake been used within the piece of media?
Medium: how is the deepfake available for viewing?
Audience visibility: how are other audience members and their reactions/discussions made available to the viewer?
Information type: what kind of information is the deepfake (i.e., is it intended to mislead)?
Representation: how is the target being represented in the deepfake?

No.	Content details					Characteristics						Notes
	Title of media	Description	Link	Year	Category	Authorization	Disclosure	Function	Medium	Audience visibility	Information type	
1	Deepfake Neighbour Wars	Parody reality TV show featuring deepfakes of several celebrities.	https://www.theguard	2023	Entertainment	Unauthorized	Yes	Novelty/spectacle	Broadcast	General public	Information	
2	Zone Out	Experimental short film created by entirely by a GAN featuring deepfake performances.	https://www.youtube.	2018	Entertainment	Authorized	Yes	Novelty/spectacle	Video sharing platform	Online audience	Information	
3	Rogue One: A Star Wars Story	Film featuring deepfake performance of deceased actor Peter Cushing.		2016	Entertainment	Target deceased	No	Role reprisal	Public release	General public	Information	Production uses CGI technique similar to deepfake technologies.
4	Star Wars: The Rise of Skywalker	Film featuring deepfake performance of deceased actor Carrie Fisher.		2019	Entertainment	Target deceased	No	Role reprisal	Public release	General public	Information	Production uses CGI technique similar to deepfake technologies.
5	Obi-Wan Kenobi	TV series featuring deepfake voice performance of retired actor James Earl Jones.		2022	Entertainment	Authorized	No	Role reprisal	Streaming service	General public	Information	
6	The Mandalorian	TV series featuring de-aged deepfake performance of actor Mark Hamill.		2021	Entertainment	Authorized	No	De-aging/role reprisal	Streaming service	General public	Information	Production uses CGI technique similar to deepfake technologies.
7	Finding Jack	Film in development set to feature deceased actor James Dean.	https://www.independ	TBC	Entertainment	Target deceased	TBC	Original performance	n/a	n/a	Information	Film remains in development but has received media attention.
8	Roadrunner: A Film About Anthony Bourdai	Documentary featuring deepfake voice performance of deceased chef Anthony Bourdain.	https://www.newyork	2021	Entertainment	Target deceased	No	Role reprisal	Streaming service	General public	Disinformation	Documentary features voice narration by Bourdain, some of which are AI-generated.
9	Harry Styles Wants More Berries	Parody video featuring deepfake of singer Harry Styles.	https://www.tiktok.co	2020	Social	Unauthorized	Yes	Novelty/spectacle	Social media	Online audience	Information	Video features a deepfake of Styles singing lines from 2019 song "Watermelon Sugar"
10	Jerry Seinfeld in Pulp Fiction	Parody video featuring deepfake of comedian Jerry Seinfeld inserted into 1994 film "Pulp Fiction".	https://www.youtube.	2021	Social	Unauthorized	Yes	Novelty/spectacle	Video sharing platform	Online audience	Information	
11	MBN News	News broadcast featuring deepfake of news anchor Kim Joo-Ha.	https://www.youtube.	2021	Journalism	Authorized	Yes	Novelty/spectacle; journalism	Broadcast	General public	Information	Deepfake anchor used for rapid reporting on MBN News online platform.
12	Mark Zuckerberg Instagram Post	Satirical video featuring deepfake of Facebook founder Mark Zuckerberg.	https://www.vice.com	2019	Social	Unauthorized	Yes	Novelty/spectacle; campaigning	Social media	Online audience	Information	
13	Volodymyr Zelenksy Surrenders	Disinformation video featuring deepfake of Ukrainian President Volodymyr Zelensky.	https://www.youtube.	2022	Political	Unauthorized	No	Deception	Social media	Online audience	Disinformation	
14	Dove Toxic Influence	Awareness campaign video featuring deepfake videos of participants giving harmful beauty advice.	https://www.youtube.	2022	Social	Authorized	No	Campaigning	Video sharing platform	Online audience	Information	
15	Dalí Lives	Museum exhibition featuring an interactive deepfake of deceased artist Salvador Dalí.	https://thedali.org/exh	2019	Educational	Deceased	Yes	Interactivity	Public exhibition	General public	Information	
16	Snoop Dogg Synthesia Ad	Advertisement for MenuLog featuring a partial deepfake of rapper Snoop Dogg to change lyrics from previous version of advertisement for JustEat.	https://www.synthesia	2020	Promotional	Authorized	No	Variation	Broadcast	General public	Disinformation	
17	JFK Unsilenced	Vocal deepfake of deceased US President John F. Kennedy delivering the speech he intended to give the day of his death in 1963.	https://www.cereproc	2018	Educational	Unauthorized	Yes	Novelty/spectacle; education	Private webpage	No audience	Information	
18	In Event of Moon Disaster	Exhibition featuring deepfakes of deceased US President Richard Nixon delivering a televised address declaring that the 1969 Apollo 11 mission ended in disaster.	https://arts.mit.edu/in	2019	Educational	Unauthorized	Yes	Novelty/spectacle; education	Public exhibition	General public	Information	
19	Oliver Taylor	Suspected deepfake image used to legitimise social media account of journalist "Oliver Taylor"	https://futurism.com/t	2020	Journalism	Authorized	No	Deception	Social media	No audience	Disinformation	
20	Celebrity deepfake pornography	Numerous instances of deepfake pornography targeting female celebrities (e.g., Gal Gadot, Emma Watson)	https://www.washingt	2017	Pornography	Unauthorized	Yes	Manipulation	Video sharing platform	Online audience	Disinformation	
21	Welcome to Chechnya	Documentary employing deep fake to alter faces of victims of torture in Chechenia while telling their experience, in order to preserve their safety and personal identity	https://scholar.google	2020	Safeguard	Authorized	Yes	Personal protection	Public release	General public	Disinformation	
22	Lola Flor for Cruzcampo	A famous deseased Andalusian artist is featured in a beer brand ad	https://www.youtube.	2021	Promotional	Authorized	Yes (?)	Spectacle/advertising	Video sharing platform	Online audience	Disinformation	
23	Hotel du Temps	Tv series featuring deceased stars and politicians who give biographical interviews	https://duckduckgo.co	2022	Entertainment	Authorized	Yes	Spectacle	Broadcast	General public	Information	
24	Dimensions in Testimony	Museum exhibition featuring deceased Shoah witnesses telling their experiences as holograms	https://duckduckgo.co	2016-2	Educational	Authorized	Yes	Education	Public exhibition	General public	Information	

25	Malaria Must Die	Awareness campaign video with deepfake video of David Beckam speaking nine different languages in nine different voices	https://news.sky.com/	2019	Educational	Authorized	Yes	Education/campaigning/ variation	Video sharing platform	Online audience	Information	
26	Tom Cruise Deep Fakes	Belgian visual effects artist Christopher Ume creates short Cruise videos with the collaboration of one of his well-known impersonators, Miles Fisher. He uploads them to Tik Tok as a joke. The videos have 10 million views.	https://www.youtube.	2021	Entertainment?	Unauthorized	No	Spectacle	Social media	Online audience	Information	
27	Portrait of Edmond Belamy	A group of artists create a portrait through GANs inspired by the inventor of GANs. The portrait is sold for $432,000.	https://www.theverge	2018	Art	Not applicable	Yes	Spectacle/art	Public exhibition	General public	Information	
28	Manus Island and Nauru	AI-generated images of Australia's offshore immigration centres based on witness statements.	https://www.theguarc	2023	Educational	Authorized	Yes	Education/art	Public exhibition	General public	Information	
29	Deepfake geography/satellite images	AI-generated satellite images of cities and landscapes could be used to deceive military strategies, suppress climate science, and hide evidence of atrocities.	https://www.engadget	2021	Political	n/a	No	Deception	n/a	n/a	Disinformation	
30	ThisClimateDoesNotExist	AI-generated images of landmarks in extreme environments to visualize the effects of climate change.	https://thisclimatedoe	2023	Educational	n/a	Yes	Education	Private webpage	Online audience	Information	
31	Vladimir Putin announcies evacuation	Deepfake broadcast of Russian president Vladimir Putin announcing the evacuation of Russian regions bordering Ukraine.	https://www.ilpost.it/	2023	Political	Unauthorized	No	Deception	Public broadcast (TV and radio)	General public	Disinformation	