AI AND DEEPFAKE TECHNOLOGY IN TIMES OF CRISIS OF TRUST IN MEDIA

Is It Really a Grave Threat to Journalism?

DESISLAVA SOTIROVA

AI И ТЕХНОЛОГИЯТА DEEPFAKE ВЪВ ВРЕМЕНА НА КРИЗА НА МЕДИЙНОТО ДОВЕРИЕ

Дали това наистина е сериозна заплаха за журналистиката?

ДЕСИСЛАВА СОТИРОВА

e-mail: d.sotirova@fjmc.uni-sofia.bg

ORCID: 0009-0009-9468-061X

DOI: https://doi.org/10.60060/MLg.2025.17.43-53

Abstract: While the concerns about media trust have historical roots, the dynamics have evolved with technological advancements, the rise of digital media, and the challenges posed by disinformation in the 21st century. Deepfake – the media content (text, images, videos) created by AI technology – is just another concern for journalists, politicians and society in the current media trust crisis. It is a tool that can be successfully applied in an information warfare. To be honest, face and voice spoofing has always existed, but deepfake technology may now make fake videos faster, cheaper, with high quality and is accessible to anybody. Although deepfake videos have the potential to undermine trust in journalism, their actual impact on audiences remains limited due to information overload, increased suspicion and skeptical attitude, decreased trust in media and advanced audience adaptability for orientation in the information flow.

Key words: deepfake technology, media trust, journalism, AI, disinformation

Резюме: Въпреки че темата за доверието към медиите има исторически корени, въпросът отново се повдига заради технологичния напредък, възхода на цифровите медии и предизвикателствата, породени от дезинформацията през XXI в. Дийпфейк – медийното съдържание (текст, изображения, видеоклипове), създадено чрез изкуствен интелект (ИИ) – е допълнително притеснение за журналисти, политици и общество в настоящата криза на доверие в медиите. Дийпфейк технологията е инструмент, който може успешно да се прилага в съвременната информационна война. Но в интерес на истината, фалшифицирането на лица и гласове винаги е съществувало чрез други програми и способи, но с дийпфейк технологията фалшивите видеоклипове се правят по-бързо, по-евтино, с високо качество и са достъпни за всеки. Въпреки че дийпфейк видеата имат потенциал да подкопаят доверието в журналистиката, тяхното реално въздействие върху аудиторията остава

ограничено поради пренасищане с информация, повишено скептично отношение към медиите, намалено доверие и адаптивност на аудиторията за ориентиране в „информационния океан“.
Ключови думи: дийпфейк, медийно доверие, журналистика, ИИ, дезинформация

The World Economic Forum's Global Risks Report 2024 ranks misinformation[1] and disinformation as the number one threat the world faces in the next two years. The prediction is based on their potential impact on reducing trust in the media and institutions. In the context of media trust crisis where 'the digitization of journalism has destabilized assessments of trust even more' (Moran, Rachel E., 2022: p.39) deepfakes amplify the crisis of trust in media, including news media. Media trust can decrease for various reasons, and these factors[2] can be interconnected. Media trust refers to the level of confidence or faith that the audience has in a particular news source or news outlets in general. It reflects the extent to which people believe that a news organization provides accurate, reliable, and unbiased information. Here should be noted that media should not be trusted blindly – but it is important the audience, the public to believe that there are journalistic standards, ethics, and accountability that allow for the separation of ethical journalism from propaganda. Trust in the media as a system should not be unreserved, but critical. People should check sources, compare points of view, recognize attempts for manipulation. Therefore, the realistic and responsible position of the modern news consumer is not faith, but informed trust.

This believe in digital age reports a decline, because of the enormous amount of unverified information, and that's why we talk about crisis of trust in

[1]In this article misinformation is used under the concept of James H. Kuklinksi – 'people are misinformed when they confidently hold wrong beliefs'. More: Kuklinski, James, Quirk, Paul, Jerit, Jennifer, Schwieder, David, Rich, Robert F. Misinformation and the Currency of Democratic Citizenship. 2000.

[2]The main factors for decreasing trust in media are: the spreading of false or misleading information, polarization and bias (political, ideological, etc.) media; lack of transparency about ownership, sources, finances, editorial politics; representing news without background (especially in news from around the world); declining trust in institutions (because of scandals, corruption, economical/social problems, etc) as whole that can spill over into a lack of trust in media; mistakes in reporting and failure to correct inaccuracies, as well as unethical behavior; relying on and believing in information from unidentified or suspicious profiles/pages in social media that deliver information that contradicts the official (often labeled as 'alternative').

media. In this context, 'with an unprecedented level of realism, speed, scale, and ability to personalize disinformation deepfakes could contribute to the broader problem of fake news on social media' (Appel, Markus; Prietzel, Fabian, 2022: p.22). The crisis of trust in media can be seen not only as a weakness of a democratic society, but also as an **adaptive mechanism** to cope with the information overload. In an environment where a huge amount of content is disseminated without fact-checking, audiences often resort to skepticism as a form of self-defense. In some parts of the world and in some countries, the crisis of trust in media is not simply a defensive reaction, but the result of a specific historical and political context. Weak governments, widespread corruption, injustice, and an ineffective judicial system contribute to a lasting erosion of trust in institutions. Because the media are often closely tied to these structures – whether through dependence on funding, political influence, or lack of transparency – they too become objects of suspicion.

In this environment of low trust in institutions and media, the emergence of deepfake technology adds another layer of complexity to the media landscape, further blurring the lines between fact and manipulation. One of the most visible and impactful uses of this technology is the creation of AI-generated videos featuring public figures, designed to deceive or provoke strong emotional reactions. Videos made by deepfake technology are usually sensational in order to make them highly shareable on social media. Such technology can be used in disinformation and misinformation campaigns to manipulate users, to influence election campaign in foreign countries or to reinforce the polarization of the population on a particular issue.

According to Merriam-Webster Dictionary deepfake is 'an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said'. In Britannica deepfake is defined as 'AI-generated synthetic media, including images, videos, and audio, generated by artificial intelligence (AI) technology that portray something that does not exist in reality or events that have never occurred'. They can be used for fun, scientific research, but also for scam. Sometimes they're used to impersonate people like politicians

or world leaders, for intentionally misleading people about what they have said. There were cases of deepfakes of journalists at American and European televisions whose faces and voices have been used to present disinformation and outright lies.

Indeed, deepfake technology is used as a tool in the Russian information warfare (Smith, Hannah; Mansted, Katherine, 2020: p.11) and not coincidentally, there were concerns in the USA for foreign interfering from both the Democratic and Republican parties because of the 2024 Presidential election[3]. The deepfake technology usually can be used by domestic or international actors for disinformation purposes such as amplifying the polarization in society, ideological conflicts and political repression. AI-generated videos and deepfake technologies only increase the speed and effectiveness of propaganda and disinformation. Measuring the impact of deepfakes is currently quite limited, not to mention how it could actually be measured – whether it is long-lasting or has a short-term effect.

Deepfakes of political figures – recent cases in the USA, the UK, the Philippines and Bulgaria

As of 2025 deepfakes[4] of politicians are mostly distributed on Telegram[5], TikTok and X. These platforms allow for rapid and wide dissemination. The viral nature of short-form and sensational content increases the likelihood that such videos will reach vast audiences before any verification or debunking can take place. This immediacy amplifies their potential impact, especially among users who consume news passively or primarily through social media. The problem is compounded by the platforms' algorithms, which tend to favor emotionally charged or controversial material, further fueling the visibility and potential

[3]During the US Presidential election in 2016 there was a controversy over the potential impact of fake news and extremely biased news on social media on Donald Trump's victory over the candidate of the Democratic Party Hillary Clinton. Therefore after the presidential race the spread of fake news on social media became a public concern in the US.

[4] Deepfake pornography images and videos are often created to offend or to 'prank', as sometimes is described, a real person whose face is superimposed onto explicit material without consent. For instance, South Korea faces deepfake porn crises and victims are often underaged girls or women.

[5]Millions of People Are Using Abusive AI 'Nudify' Bots on Telegram. Available from: https://www.wired.com/story/ai-deepfake-nudify-bots-telegram/.

influence of deepfake videos. Deepfakes featuring political figures have increasingly appeared in recent years, often timed to coincide with elections, political crises, or major international events, which heightens their potential to mislead and manipulate public opinion.

In November 2023 a video appeared on social media where the London Mayor Sadiq Khan[6] is supposed to have said inflammatory remarks before Armistice Day on 11 November 2023[7]. He claimed that this video almost caused 'serious disorder'. The intelligence established that the clip used AI to create a replica of Mr. Khan's voice. The fake audio is calling for Armistice Day to be re-scheduled for a pro-Palestinian march[8]. The words scripted are: 'What's important and paramount is the one-million-man Palestinian march takes place on Saturday'. In the video Mr. Khan's voice is imitated saying: 'I control the Met Police, they will do as the Mayor of London tells them' and 'the British public need to get a grip'. Although the Metropolitan Police confirmed that the recording was fake and did not constitute a crime, the incident caused serious public reactions.

In December 2023 a video with the President of Republic of Bulgaria Rumen Radev occurred on social media suggesting people to invest in Lukoil company. A recording of an official event attended by the Bulgarian president was used. A supposed 'reporter' asks the president to 'tell about the Lukoil project and why he made such a decision'. Then the manipulated voice of the President is heard saying: 'I have been considering the implementation of this project for a long time, because many of our citizens are always looking for projects in which they can profitably invest in order to receive additional passive income'. According to the information of the investigating authorities the purpose was a financial scam.

This is not the first deepfake video with the face of a Bulgarian politician. In October 2023, the press service of the Council of Ministers warned that a

[6]Sir Sadiq Aman Khan is the first Muslim Mayor of London and has been serving on this position since 2016. He is also the first Mayor of London who wins a third term.

[7]The deepfake video emerged during an already-tense political row. The first Hindu Prime Minister of the United Kingdom Rishi Sunak is faked to have said the pro-Palestinian marches in a different part of central London were 'disrespectful' on Armistice Day. Armistice Day marks the moment when World War One ended on 11 November 1918.

[8] Pro-Palestinian demonstrations in central London are a common sight since the start of Israeli rocket fire on the Gaza Strip. Such rallies are still organized.

deepfake clip circulated on social media using the image and voice of the then Prime Minister Nikolay Denkov for a fraud. Bulgarian journalists' identities have also been used both to spread disinformation and various fraudulent schemes.

In late April 2024 an audio of the Philippine President Ferdinand Marcos Jr. appeared on social media calling for a military action against the People's Republic of China. The President's communications office quickly denied the recording was authentic and confirmed the audio was fake. The manipulated audio recording emerged amid increasingly aggressive actions by the People's Republic of China (PRC) against the Philippines in a part of the South China Sea. Beijing claims as its territory, despite a 2016 international tribunal ruling that the region is within the Philippines' exclusive economic zone. The current issue with China are the Philippines ships in disputed waters where encounters with China have become more frequent. Marcos had repeatedly said he's not trying to provoke Beijing but asserts its nation rights. The audio, where the President Marcos is heard to be urging his military to intervene if China poses a threat to the Philippines, was quickly deleted.

In May 2024 a deepfake of the United State Department spokesman Matthew Miller also appeared on social media, after Shift on Ukraine Attacks in Russia. It happened a day after U.S. officials said Ukraine could use American weapons in limited strikes inside Russia. The fabricated video was created from real footage from press briefings. It shows Matthew Miller suggesting the Russian city of Belgorod was a legitimate target of Ukrainian strikes. Belgorod is located near the Russian - Ukrainian border, just 25 miles north of Ukraine's border with Russia. This claim was spread across Russian media and social media, including Telegram and X (formerly Twitter), and was used by Russian propagandists, including Dmitry Rogozin, the former head of Roscosmos, to justify Russia's attacks on Ukraine and undermine support for Kiev. The video was shared also by the Russian President Vladimir Putin's adviser and Chairman of Russia's Human Rights Council Valery Fadeyev. He shared the deepfake video on his Telegram channel and later deleted it.

All these examples show that deepfake attacks are a global phenomenon. They are not isolated cases, but part of a broader information strategy with a

simple goal: influence public opinion, provoke panic or confusion, and encourage distrust. Although these cases receive widespread media attention and provoke reactions from official institutions and society, they did not lead to policy shifts, public unrest, or measurable changes in trust. In many cases, deepfakes do not initiate crises, but simply reflect or amplify existing divisions. Their power lies more in symbolic disruption than in creating tangible real-world consequences.

Deepfake – a symptom of the crisis of trust in media

Deepfake technology should be seen as a symptom of an already existing and deepening crisis of trust in modern media and institutions (Popa, Claudiu, et.al, 2025). Its emergence and increasing use exploit the lack of trust in media. One of the best tactics for coping with deepfakes is the fast institutional reaction and fact-checking (Rucinska, Silvia, et.al., 2022). The quick refutations certainly play a key role, but they do not exhaust the impact of deepfake videos on society because usually the first message people see or hear often leaves the strongest mark on their minds. But do they actually have an impact with consequences in reality? The harms of deepfake technology in social media should be analyzed in the broader context of public fatigue, crisis of trust in institutions and media, and information overload. The impact of the deepfake technology has yet to be studied, as it is very likely that its influence on the audience is not universal – it depends on the specific audience: age group, political attitudes, education, media literacy, and ability to check facts.

The claim that deepfake is a symptom rather than a cause of the crisis of trust in the media is supported by the following arguments, which reveal the deeper systemic factors underlying the erosion of trust.

- The crisis of trust in media has started even before the rise of deepfake on social media. This phenomenon is another stage in the process, but not the original destroyer of authority. An audience exposed to constant propaganda and manipulation develops ‘information immunity’ or apathy.
- People begin to doubt everything, including disinformation, which reduces its impact but also leads to withdrawal from participation in public debate – i.e. deep, passive lack of trust.

- Technology is most effective when it reinforces already existing polarizations by serving as 'visual evidence' of political bias (Twomey, John, et al., 2023).
- Protection against deepfake should be not only by technical ways, but also editorial and institutional – new realities require new codes and verification skills.

In a context where audiences are increasingly being asked to verify the authenticity of information and the integrity of sources, deepfake content finds itself in fertile ground for dissemination. It is a reflection of a cultural and media landscape in which the boundaries between reality and manipulation have become blurred, and truth has lost its authority.

Challenges to the verification processes of Deepfakes

In response to the rise of deepfakes and their destabilizing effects, both the media industry and technology sector are exploring strategies to **detect and counter AI-generated content**. AI-powered tools are being developed to identify digital inconsistencies, watermark original footage, and trace manipulated videos. However, the technological race between creators of deepfakes and their detectors remains ongoing, with each side rapidly evolving. This technological arms race highlights the need for **a multi-layered response** that goes beyond detection and includes **transparent editorial practices, stronger verification protocols, and international cooperation** in combating malicious content.

At the same time, **media and journalists play a crucial role** in defending the public against manipulated narratives. Strengthening investigative journalism and reinforcing fact-checking efforts are key to rebuilding trust. Media outlets must also become more transparent about their sources, editorial choices, and the context in which news is reported—particularly in an era where visual evidence can be faked.

Equally important is the role of education and media literacy, especially among younger generations who consume most of their news online. Citizens must be equipped not only to recognize deepfakes and disinformation, but also to critically evaluate the motives behind the content they encounter. Empowering

people with tools to question, verify, and interpret information is one of the most sustainable strategies for countering the influence of synthetic media and restoring trust in authentic journalism. The digital era is marked by the unprecedented dissemination of misinformation across multiple platforms, but it is equally defined by our capacity to reaffirm truth through diverse and far-reaching means.

We've already seen AI created images and videos of non-existing people in reality on Instagram but they are labeled as AI-generated. Along with videos with artificial faces there are Reels on Instagram, Facebook and TikTok which are not noted as being generated by artificial intelligence but present a false reality. Such as Reels showing big cities taken over rats.

And what if the content we see on social media or even on TV news is so artfully fabricated that it cannot be recognized by journalists, nor by the audience as fake? And who will be responsible for spreading disinformation? Of course, there are applications developed for detection of AI-generated media content. They involve the deployment of a high-accuracy deep learning model trained on a diverse dataset, including authentic and synthetic images, with a focus on content generated by advanced AI tools such as diffusion models. In the near future journalists will have difficulties to check if a video is fake or not without advanced tools and this may mean delaying in reporting. Hence, newsrooms will need to invest in software, apps, trainings, and teams to verify videos and images taken from another source or person.

Still deepfake technology used in a video can be recognized because it seems like a filter on it. When it comes to AI-manipulated media, there is no single sign of fake content. However, there are a few deepfakes features to recognize them (Detect DeepFakes:2021). But let's say in two years' time that obvious difference could be fixed, that's why people should begin to train themselves 'to be more aware of fake images and videos, especially when the videos are asking to send money or personal information, or making outrageous claims that seem unusual for the person who appears to be making them' (University of Virginia).

In conclusion, the rapid development of AI-generated media content presents both opportunities and significant challenges for journalists and society as a whole. While deepfake detection tools and technologies are evolving, they

may not always keep up with the complexity of AI-generated media content. This leaves both journalists and audiences vulnerable to disinformation. Despite the technological sophistication and potential impact of deepfake videos, their actual effect on public attitudes is often smaller than expected. The reason for this is rooted not so much in the power of technology itself, but in the state of the society. Many video consumers already approach everything they see and hear with skepticism – not only because they recognize the possibility of manipulation, but also because trust in the media, institutions, and public discourse has been deeply eroded. In conditions of information overload, people often become apathetic, cynical, or simply indifferent – including to obvious fakes. It is precisely this 'trembling' that acts as a vaccine against manipulation, but also creates dangerous ground for political apathy.

But here comes the question of journalists' responsibilities. Who is responsible for sharing media content created by deepfake technology if applications for detection don't recognize that the video or image is fake? The journalist, the media or the creator of the application? Ultimately, the responsibility to tackle the spread of AI-generated disinformation cannot rest solely on journalists. The media, AI developers, regulators, policymakers and even individual users all play a role in ensuring the integrity of information. The solution does not lie solely in algorithms or regulations, but in the **collective responsibility** of journalists, educators, policymakers, and citizens to defend truth, demand transparency, and think critically. In a world where deception can go viral, **truth must be made visible, verifiable, and resilient**. Only through collaboration and transparency can we hope to defend truth in times where fiction can be digitally manufactured to appear more real than reality itself.

Bibliography:

Appel, Markus, Prietzel, Fabian (2022) The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4). Available from: https://academic.oup.com/jcmc/article/27/4/zmac008/6650406 [Accessed 17 Jan. 2025].

Kuklinski, James, Quirk, Paul, Jerit, Jennifer, Schwieder, David, Rich, Robert F. (2000) Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3). Chicago: The University of Chicago Press.

Moran, Rachel E. (2022) The so-called "crisis" of trust in journalism. In: Allan, Stuart ed. *The Routledge Companion to News and Journalism*. 2nd ed. London: Routledge.

Popa, Claudiu et al. (2025) Deepfake technology unveiled: The commoditization of AI and its impact on digital trust. Available from: https://arxiv.org/pdf/2506.07363 [Accessed 9 Jun. 2025].

Rucinska, Silvia, Fecko, Miroslav, Mital, Ondrej (2022) The role of public authorities in responding to disinformation. In: *Central and Eastern European eDem and eGov Days (CEEeGov)*. ACM. Available from: https://dl.acm.org/doi/abs/10.1145/3551504.3552297 [Accessed 9 Jun. 2025].

Smith, Hannah, Mansted, Katherine (2020) Weaponised deep fakes: National security and democracy. *Australian Strategic Policy Institute (Policy Brief)*. Available from: https://www.jstor.org/stable/resrep25129.7?seq=3 [Accessed 19 Jan. 2025].

Twomey, John, Ching, Didier, Aylett, Matthew Peter, Quayle, Michael, Linehan, Conor, Murphy, Gillian (2023) Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. PLOS ONE, 18(10). Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0291668 [Accessed 9 Jun. 2025].

References:

BBC (2025) Thousands at pro-Palestinian rally in London. Available at: https://www.bbc.com/news/articles/cz0l34kpv51o [Accessed 15 Jan. 2025].

Britannica, 2025. Deepfake. Available at: https://www.britannica.com/technology/deepfake [Accessed 11 Jan. 2025].

BNR (2025) Bulgarian president victim of video hoax. Available at: https://bnr.bg/en/post/101919858/bulgarian-president-victim-of-video-hoax [Accessed 15 Jan. 2025].

The New York Times (2024) Deepfake of U.S. Official Appears After Shift on Ukraine Attacks in Russia. Available at: https://www.nytimes.com/2024/05/31/us/politics/deepfake-us-official-russia.html [Accessed 15 Jan. 2025].

MIT Media Lab (2025) Detect DeepFakes: How to counteract misinformation created by AI. Available at: https://www.media.mit.edu/projects/detect-fakes/overview/ [Accessed 11 Jun. 2025].

Instagram, 2025. Good morning London. Available at: https://www.instagram.com/numanuk/reel/DEe5-JRMIuR/ [Accessed 15 Jan. 2025].

Politico (2025) London's Sadiq Khan shaken by pro-Palestine deepfake. Available at: https://www.politico.eu/article/mayor-of-london-sadiq-khan-shaken-by-pro-palestine-deepfake/ [Accessed 15 Jan. 2025].

Merriam-Webster (2025) Deepfake. Available at: https://www.merriam-webster.com/dictionary/deepfake [Accessed 12 Jan. 2025].

Wired, 2025. Millions of People Are Using Abusive AI 'Nudify' Bots on Telegram. Available at: https://www.wired.com/story/ai-deepfake-nudify-bots-telegram/ [Accessed 15 Jan. 2025].

World Economic Forum (2024) The Global Risks Report 2024. Available at: https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf [Accessed 10 Jan. 2025].

Voice of America (VOA) (2025) The Matthew Miller deepfake has the attributes of Russia's information warfare. Available at: https://www.voanews.com/a/7654403.html [Accessed 15 Jan. 2025].

University of Virginia (2025) What the heck is a deepfake? Available at: https://security.virginia.edu/deepfakes [Accessed 20 Jan. 2025].

Desislava Sotirova, Ph.D. is a Chief Assistant Professor at Radio and TV Department of the Faculty of Journalism and Mass Communication, Sofia University 'St. Kliment Ohridski'. Her research interests are in the field of television communication, international journalism and media studies.

Д-р Десислава Сотирова е главен асистент в катедра „Радио и телевизия" във Факултета по журналистика и масова комуникация, Софийски университет „Св. Климент Охридски". Научните ѝ интереси са в сферата на телевизионната комуникация, международната журналистика и медийните изследвания.