MEDIATIZATION STUDIES 9/2025 DOI: 10.17951/ms.2025.9.97-115

SOFIIA VAVRIN

Maria Curie-Skłodowska University in Lublin

sophievavrin@gmail.com

https://orcid.org/0009-0001-1753-4047

Artificial Intelligence and Disinformation in Public Relations: Challenges and Countermeasures

Abstract: The rapid development of artificial intelligence (AI) has increased the threat of disinformation and deepfakes, significantly affecting public trust, political stability, and media credibility. Despite growing academic interest in AI's benefits, a research gap remains in analyzing its manipulative potential in modern public relations. This study examines how AI-generated content contributes to the spread of fake news and deepfakes and evaluates the effectiveness of existing legal, technological, and educational countermeasures. It also aims to examine the extent to which AI-generated content impacts public trust, political discourse, and media credibility. The research questions include: How does AI contribute to disinformation in PR? What legal and technical solutions exist across different countries? The analysis reveals that AI is intensifying information manipulation, while detection tools and regulations remain fragmented. A multi-level approach combining legal frameworks, education, and international cooperation is essential to address AI-driven disinformation effectively and protect public communication.

Keywords: deepfake; fake news; artificial intelligence (AI); disinformation; public relations

Introduction

Artificial intelligence (AI) is quickly becoming an important part of everyday life, radically transforming various fields of activity, from medicine to entertainment. However, along with its many benefits, new threats are also emerging. One of the most serious dangers is the use of AI tools to create disinformation and spread disinformation. This paper will discuss the nature of deepfakes, their threats, examples of disinformation generated by AI, and possible strategies to combat these problems.

In order to understand the nature of AI, firstly it is important to clarify the meaning of this term. In general, the development of AI dates back to the mid-20th century and

is associated with Alan Turing's (1950) test,[1] which allowed him to determine whether a machine is capable of thinking like a human. Another well-known definition of AI was proposed by John McCarthy. In 2007, he defined AI as follows: "Artificial intelligence is the science and engineering aimed at creating intelligent machines. Intelligence is the computational aspect of the ability to achieve goals in the real world" (McCarthy, 2007, pp. 2–5).

Different countries and organizations provide their own definitions of AI, focusing on different aspects of this technology, reflecting its diversity and complexity. By the European Parliament, AI is defined as any tool used by a program to simulate human behaviour, including thinking, planning, and creativity. This definition can be expanded, as AI is already capable of surpassing human capabilities in certain areas (European Parliament, 2020).

In the United States, AI is defined in multiple federal statutes, reflecting both technical and functional perspectives. According to 15 U.S. Code § 9401, AI is "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments" (Legal Information Institute. 15 U.S. Code). Additional notes in 10 U.S. Code § 2358 expand this definition to encompass artificial systems that can perform tasks under varying and unpredictable circumstances without significant human oversight, learn from experience to improve performance, act or think like a human, and employ techniques such as machine learning, cognitive architectures, and neural networks to achieve goals through perception, planning, reasoning, learning, communication, decision-making, and action (Legal Information Institute. 10 U.S. Code).

The Polish Ministry of Digital Affairs published the Artificial Intelligence (AI Act) Regulation on July 12, 2024. According to its definition, an AI system is a tool that uses inference techniques such as machine learning, logic and knowledge to generate results in the form of predictions, content, recommendations or decisions that can affect physical and virtual environments. AI systems operate with varying levels of autonomy, can adapt over time, and function both as stand-alone solutions and as components of other products (Ministerstwo Cyfryzacji, 2025).

In December 2020, the Cabinet of Ministers of Ukraine approved the Concept of Artificial Intelligence Development in Ukraine. The Concept uses the term in the following meaning: "Artificial intelligence – is an organized set of information technologies that can be used to perform complex tasks through the use of a system of scientific research methods and algorithms for processing information received or independently created

[1] In 1950, Alan Turing proposed a test he called the "simulation game". Based on it, AI can be defined as follows: "Artificial intelligence" is any computer that successfully passes the Turing test."The Turing Test" is a game involving three people: (1) a human, (2) a computer, and (3) a human judge. The judge does not see the other participants and can only communicate with them via text. If the judge is unable to reliably determine which of the participants is a machine, the computer is considered to have passed the test (Turing, 1950, pp. 433–460).

during work, as well as creating and using its own knowledge bases, decision-making models, algorithms for working with information and determining ways to achieve the tasks" (Rozporiadzhennia Kabinetu Ministriv Ukrainy, 2020).

The European Parliament emphasizes the imitation of human behavior, including thinking and creativity, while the definitions of Ukraine and Poland focus on technical aspects such as algorithms and adaptability. Ukrainian law emphasizes the ability of AI to create knowledge on its own, while the Polish one focuses on the autonomy of systems and their impact on the physical and digital environment. In the United States, AI definitions highlight both functional and operational dimensions, emphasizing machine-based systems capable of making predictions, recommendations, or decisions under varying and unpredictable circumstances. Despite these differences, all definitions recognize the ability of AI to perform complex tasks and adapt over time.

Given these conceptual differences across jurisdictions, it becomes essential to analyse how AI-driven disinformation is addressed in practice and what methodological tools can be used to examine these challenges effectively. To address this problem, the study employs a qualitative approach combining a literature review, case study analysis and comparative legal analysis. High-profile examples of AI disinformation (e.g. the fake videos of Narendra Modi, Nancy Pelosi and Volodymyr Zelenskyi) were selected to illustrate practical consequences for political communication and public trust. Legal frameworks across countries – including the EU AI Act, the U.S. No Fakes Act (2025; revised draft), and regulations in China, Singapore and South Korea – were compared to evaluate the diversity of countermeasures. Secondary sources such as peer-reviewed articles, policy papers and legislative documents formed the basis of analysis, which was conducted through thematic coding. While this design provides robust insights, its reliance on secondary data highlights the need for future research involving interviews or surveys with PR practitioners and policymakers. The study formulates two main research questions: RQ1: How does AI contribute to disinformation in political PR? RQ2: What legal and technical solutions exist across different countries? Building on the identified research gap, it is important to consider how AI technologies are practically applied to generate and disseminate disinformation in the public sphere.

Literature review

Existing literature extensively explores the capabilities of AI in automating media processes, generating content, and improving data analysis within the field of public relations. These issues are discussed, among others, by Blankenship (2021) and by Mehan (2024). Scholars such as Pepping et al. (2021) or Lyon and Tora (2023) have also highlighted the ethical risks posed by AI-generated disinformation and its potential to disrupt democratic processes. Deepfake technology, in particular, has received growing academic attention due to its ability to manipulate audiovisual content, as noted

by Buzzfeed and NPR, shaping public perception in new ways. While many studies have addressed AI's technical mechanisms or its societal implications separately, few have examined how these technologies intersect specifically with the strategic field of public relations and its role in shaping public trust.

Moreover, although legal frameworks such as the EU's AI Act (21 May 2024) and the US No Fakes Act (2025; revised draft) have been analyzed in regulatory literature, comparative studies assessing their effectiveness in the context of PR-driven disinformation remain limited. Similarly, existing research tends to focus on individual national approaches rather than providing a cross-border perspective on legal and technical countermeasures. These findings highlight a critical gap: while the technological capabilities and regulatory responses of AI-generated disinformation are relatively well-documented, less attention has been paid to its practical implications within the strategic processes of public relations. In particular, there is a need to understand how AI technologies influence public perception, media credibility, and the strategic management of information in PR contexts.

The manipulative potential of AI tools in public relations can be further contextualized by applying established communication theories. One of the most relevant frameworks is agenda-setting theory, which argues that the media do not tell people what to think, but rather what to think about (Kuzhman, 2016, pp. 274–276). In the age of AI, algorithmically generated disinformation can artificially elevate specific issues, ensuring that fabricated narratives receive disproportionate attention while other topics are overshadowed. This mechanism directly illustrates how AI can distort public agendas by amplifying false content through bots, fake accounts, and algorithm-driven virality. This mechanism involves three key steps: 1) AI identifies topics likely to trigger engagement; 2) automated accounts and bots disseminate fabricated content; 3) platform algorithms detect high engagement and further boost visibility, creating a feedback loop that amplifies false narratives. Such processes can significantly distort public attention and perception, what illustrates how AI-driven content manipulates the public agenda (Kumar & Shah, 2018, pp. 1–4).

Closely related framing theory focuses on how the presentation of information influences public interpretation and attitudes (Goffman, 1986, pp. 1–5). The AI-generated deepfakes and synthetic news reports act as powerful framing devices: they create alternative realities, shift emotional resonance, and guide audiences toward particular evaluations of political actors or events.This mechanism works by manipulating visual, auditory, and contextual cues to make fabricated content appear authentic. Subtle changes in facial expressions, gestures, or vocal intonation can evoke specific emotions such as trust, fear, or outrage, which in turn influence how audiences interpret the message (Qiao & Zhou, 2024). For example, manipulated videos of public figures do not simply spread false facts but also reframe political discourse in a way that undermines trust in legitimate institutions. Detailed examples of such deepfakes are examined in the subsequent "Case Studies" subsection.

The integration of these frameworks into the analysis highlights that the threats posed by AI are not limited to technological sophistication. Instead, they represent a strategic communication challenge: AI reshapes the processes of agenda-setting and framing, thereby destabilizing the credibility of media and public institutions. Despite the applicability of such theories, comprehensive studies explicitly linking agenda-setting and framing to AI-driven disinformation in PR remain scarce.

Case studies: Using AI to create disinformation

In today's international information relations, the impact of disinformation on public opinion, political processes, and state stability is becoming increasingly noticeable worldwide. Instead of traditional methods such as military pressure or economic sanctions, disinformation is used as an effective tool to influence public opinion and political processes. It is capable of shaping false opinions, interfering with democratic mechanisms, and undermining trust in government institutions (Marchuk, 2024, pp. 233–236).

The use of AI algorithms to produce fake news automatically, manipulate images and videos, and generate realistic audio files that imitate the voices of famous people, makes it much more difficult to identify true and false information. This becomes particularly dangerous when such materials are used to manipulate electoral processes, undermine trust in state institutions, or foment social and ethnic conflicts.

Currently, there are more than 10,000 AI-related projects in the world. The most popular and widespread of them are presented in Table 1. The tools listed in this table can be used to create fake news, altered images, and videos to spread disinformation, provoke conflict, and propaganda.

Table 1. The most popular tools of generative AI in 2025

The field of AI application	Programs with AI	The field of AI application	Programs with AI
Voice/sound	HeyGen Underduck Revoicer Voicify Fliki Eleven labs	Create mockups	Figma WiXADI Canva Microsoft Designer Uizard VisualEyes Adobe Express DESIGNS.AI
Music	Mubert Soundful Audiocraft Aiva	Productivity	Fathom Reclaimai COGRAM Clara Otter.ai

The field of AI application	Programs with AI	The field of AI application	Programs with AI
Image	Midjourney Dalle Bing Canva Adobe Firefly Stable Diffusion Leonardo.AI STOCKIMG.AI Hotpot Jasper	Creating videos	PICTORY PIKA Midjourney Stable diffusion Synthesia HeyGen Runway Ebsynth TOPAZ LABS
Self-care	FITBOD Sleep.ai Headspace MealMate	Video editing	Descript Topaz Video Ai Visla
Education	Kena.AI birdbrain Babbel Moises Duolingo Max	Creating animation	DEEPMOTION Cascadeur
General topics/dialogue	Grok, Perplexity, ChatGPT, Bard, MPT-7B, Chatsonic, Claude.ai, KoalaWriter, ProWritingAid, Grammarly, Notion AI, INGER, Smartwriter.ai, Phrase, jasper, adcopy, WRITER + DeepSeek (2025)		

Source: (Solis, 2023).

AI algorithms are able to spread disinformation on social media in a targeted manner. They achieve this by adapting content to the interests and beliefs of specific audiences, which greatly increases its impact and effectiveness. This process amplifies the reach and impact of fake news, promotes distrust and disorientation among the population, and fuels conflicts and misunderstandings between states. Modern technologies for creating fake visual materials are one of the most dangerous tools of disinformation. According to statistics, 65% of people perceive and remember visual content better than textual information (Soyiba et al., 2019). Users can see an image with accompanying text on blogs but do not read the full article. This is problematic because the article's content can significantly alter their understanding of the situation, especially when such materials are rapidly distributed without verification.

The example of Prime Minister of India, shown in Image 1, demonstrates how quickly people react to and share visual information without verifying its authenticity. In this case, the dissemination of the image caused reputational damage to both the Prime Minister and his party.

Image 1. "Fake photo" published in India in April 2019 (1 original photo, 2 fake photo)
Source: (The Times of India, 2019).

The image featured a fake photo of the Indian Prime Minister allegedly touching the feet of Congress President Sonia Gandhi, when the real image shows him paying respect to Lal Krishna Advani, the former leader of the political party. Despite the clear signs of editing, Indian citizens quickly believed the falsification, which amplified the negative media effect and affected the credibility of Modi and his party (The Times of India, 2019). Not only do fake photos pose a particular threat today, but also more sophisticated manipulations with video and audio, where politicians can "say" things in their own voice that they did not actually say. Such technologies make disinformation even more convincing, as society is more inclined to trust audiovisual materials.

Disinformation by foreign states and related non-state actors is regularly presented as the main threat to Western democracies and the international institutions they have created. Awareness of the danger of manipulating information for political purposes has increased dramatically after repeated external interference in the internal political process of Western countries. The high-profile cases of such interference took place during the US presidential election (2016), Brexit process (2016), the referendum in the Netherlands on the Association Agreement with Ukraine (2016) and an attempt to interfere in the presidential elections in France in 2017 (Trittin-Ulbrich et al., 2020, pp. 8–25).

An interesting example of the threat caused by manipulation of photo and video materials, as well as their fast spread, was the fake videos of the Speaker of the US House of Representatives Nancy Pelosi from 2019. In those videos she is shown stumbling and slurring her speech, giving the impression that she is intoxicated (Image 2). These fakes emerged against the backdrop of President Trump's confrontation with the leader of Democratic Party. Due to their widespread dissemination on social media,

they had a significant impact on both the domestic political situation in the United States and international attitudes toward American politics (New York Times, 2019).

Image 2. Screen from the original and fabricated video with the Speaker of the House Nancy Pelosi in 2019

Source: (The New York Times, 2019).

In addition to direct security threats, such technologies undermine trust in information in general. In a world where any video can be manipulated, people begin to doubt even the truthful materials, which blurs the line between reality and fiction. This not only escalates political confrontation, but also makes it harder to fight disinformation.

One of the most dangerous tools in this area is the so-called deepfakes, a technology that allows for the falsification of video and audio by changing a person's words or actions so realistically that it becomes extremely difficult to expose the falsification. Given their growing role in shaping public perception, it is essential to examine in greater detail what deepfakes are and how they are applied in the context of public relations.

Case studies: What are deepfakes and how are they used in public relations?

Deepfake (a combination of deep learning and fake) is a digital falsification created with the help of AI, mostly in the form of faked photo and video that are so realistic that they are difficult to distinguish from the original (Westerlund, 2019, pp. 39–40). This technology works on the basis of special algorithms that analyze input data from various sources (photos, videos) and reproduce the smallest physiological features of a person – his or her facial expressions, gestures, movements in different angles and lighting. Programs such as DeepFaceLab (which covers 95% of the market), Reface, Zao, FaceApp, GauGAN synchronize these elements to create a so-called "mask" that

is superimposed on the image or video of another person. As a result, a coherent but completely fake content is formed from separate real fragments, which can be used for manipulation and disinformation (Okhrymovych, 2024, p. 47). In general, this technology is actively used in the film industry to correct shots, replace actors, or correct mistakes during filming. Unfortunately, deepfake is also used in various kinds of misleading schemes. Most often the targets of such fakes are famous personalities, influential politicians and state leaders. In addition, political technology experts point out that deepfakes are used to interfere in elections, increase political tension, manipulate public opinion and other purposes (Konoplianok, 2024, pp. 40–47).

One of the most famous cases of deepfake technology in the United States is Barack Obama's 2018 "speech". This video was created with the help of AI as part of a project organized by the University of Washington together with Buzzfeed to demonstrate the potential threats of such fakes. The video used Obama's real voice, to which synthesized lip movements were added to create the impression that he was saying the words programmed by the researchers (Silverman, 2018).

One of the most famous cases of deepfake technology being used against Ukrainian politicians is related to a fake video of Ukrainian President Volodymyr Zelenskyy calling on the nation to capitulate in the war with Russia (Image 3). This video was created at the beginning of Russia's full-scale invasion of Ukraine in 2022 with the aim of sowing panic and uncertainty among Ukrainians and the armed forces.

Image 3. Real photos of President Zelenskyy vs. fake screenshot
Source: (*Artificial Intelligence Index Report*, 2023).

The video was quickly recognized as fake, and Zelenskyy himself promptly released a genuine address in which he denied the rumors and called for unity and resistance

to aggression (Allyn, 2022). This incident highlights the importance of a critical approach to information disseminated on the Internet and the need to verify sources before sharing news. It also shows how AI technologies can be used in information warfare to destabilize society and undermine trust in governmental structures.

Another example of manipulating public opinion and attempting to discredit an opponent is the situation when Donald Trump accused Kamala Harris of using AI technologies to falsify the number of people attending her rally in Michigan. This shows how political controversies can use the latest technologies to spread false information to influence voters' emotions and perceptions. Such actions increase polarization in society and cause distrust in the political process (Horton et al., 2024).

All analyzed examples demonstrate the power of AI to create believable videos that can be used for political manipulation or discrediting public figures. This poses new challenges in the process of verifying information and combating disinformation in the digital age.

In the 21st century, we have to face these new challenges and threats. This is why we have to answer a very important and difficult question: How can we differentiate a fake from a real video? In attempting to answer this question, it is important to be mindful and pay particular attention to the following details:

- skin color – may not look like the real one;
- the edges of the mask around the face; sometimes videos are imperfect, so you may notice obvious imperfections in the reproduction of a real person;
- occlusion of the face distorts objects or covers them;
- blurred face; if you look closely, the character's face may be blurred, and this is not related to the video quality;
- light flickering; sometimes algorithms cannot fully "read" and reproduce a person's image. These secondary flickers should be visible on the screen;
- different focal length;
- inconsistency of the image[2] (Somova, 2022).

Deepfakes represent a new frontier of misinformation in public relations, creating highly realistic but fabricated audio and video content that can manipulate perceptions and spread false narratives. In response, PR professionals must adopt a proactive approach to communication. They craft clear and accurate narratives before false content gains traction, continuously monitor media and social platforms to detect and correct misleading material, and collaborate with fact-checking organizations to ensure accuracy. During crises, PR teams provide swift, evidence-based responses through official channels, press releases, and social media to maintain public trust. Leveraging credible media outlets and influencers helps amplify truthful informa-

[2] To check whether a video was generated by AI, you can use Deepware Scanner. You can also use Truly Media, a platform that includes AI-powered features to verify digital content and detect disinformation.

tion, while educational initiatives teach audiences to recognize and verify sources. Furthermore, PR specialists collaborate with stakeholders – including government bodies, NGOs, business associations, and online platforms – to support reliable news and fact-checking initiatives. Through this combination of proactive communication, monitoring, education, and collaboration, PR not only manages reputations but also protects the truth and strengthens public resilience against deepfakes and other forms of misinformation (K2 Communications, 2024).

In this regard, it is important for states and international organizations to develop complex strategies to detect and prevent disinformation based on AI technologies. This requires investments in the development of technologies to detect deepfakes, the creation of legal and ethical standards to regulate the use of AI in the information space, as well as active international cooperation in the field of cybersecurity and the fight against cybercrime. Only through joint efforts can we reduce the risks associated with the use of AI to spread disinformation and ensure the protection of international peace and security.

Building on this broader perspective, it is also important to examine how individual states are responding to these challenges through legislation and regulation.

Legal analysis: legal strategies for combating disinformation and fake news

A variety of measures to limit the spread of fake news and deepfakes have been introduced by different states around the globe. In this context, it is worth considering how different legal systems address the problem at the legislative level. In the United States, the fight against deepfakes is being conducted at both the federal and state levels. One of the newest legislative initiatives is The Nurture Originals, Foster Art, and Keep Entertainment Safe Act of 2023 (Congress.gov., 2024) and No Fakes Act (2025; revised draft). This Act is aimed at preventing the creation of digital copies of individuals without their consent or without the permission of the rights holders. At the same time, the document provides exceptions for certain areas, including news reports, public discussions, sports broadcasts, as well as documentary and biographical works. Also, parody, satire, and criticism are not subjects to the ban. The Act has received wide support among content creators, as it strikes a balance between the rights of individuals and freedom of creative expression.

Some states also have laws aimed at combating deepfakes. For example, California passed AB-730 (California Legislative Information, 2019), which bans the distribution of fabricated audio and video materials about candidates without clearly disclosing their artificial nature within 60 days before the election. And in Texas, SB-751 (2019) (LegiScan) came into force, prohibiting the intentional creation and distribution of deepfakes to influence elections.

Laws such as SB-751 in Texas and The No Fakes Act (2025; revised draft) demonstrate a growing awareness of the threats posed by deepfakes, especially in the con-

text of the electoral process and digital rights protection. The No Fakes Act (2025; revised draft) is designed to protect intellectual property and personal data by setting clear restrictions on the creation of digital copies without the consent of the relevant individuals. At the same time, the draft law provides exceptions for journalism, documentaries, and creative works, which helps to maintain a balance between human rights protection and freedom of expression.

In the European Union, the Artificial Intelligence Act (AI Act) is the key regulatory act governing the use of AI. According to the Act, a deepfake is "an image, audio or video content created or modified by artificial intelligence that imitates real persons, objects, places, organizations or events, misleading a person as to its authenticity". The AI Act introduces new requirements for deepfakes, requiring AI developers to label them in a machine-readable format to facilitate identification. Users distributing such content must also clearly indicate its artificial origin. Exceptions are provided for artistic, satirical, and law enforcement materials. The main purpose of these rules is to minimize the risks of disinformation, increase transparency, and protect citizens from manipulation. In addition to labeling content, AI providers should implement reliable technical mechanisms to help users easily distinguish between deepfakes and real content. Such solutions can include watermarks, metadata or cryptographic methods that confirm the authenticity of materials. It is important that these technologies are effective and do not create an excessive financial burden on developers. Users who distribute or publish content generated by AI must also clearly indicate its artificial origin. This rule applies not only to large companies, but also to individuals who may create deepfakes for personal purposes. At the same time, the law provides exceptions for content used in criminal investigations, art or satire. The AI Act also obliges organizations to develop codes of conduct to comply with the new requirements. The European Commission can approve such codes or define general principles for their implementation. All of these measures are aimed at effectively combating disinformation and strengthening trust in the EU information space (Eitren, 2024).

The Digital Services Act (DSA, 2023) plays a key role in the fight against disinformation, which is based on the principle that "what is illegal offline should be illegal online". In other words, the rules that apply in the physical world should also apply in the digital space. This law demonstrates the EU's desire to create a safe digital environment in which the fundamental rights of users are guaranteed. The reason for the DSA was the growing use of online services to manipulate algorithms to spread disinformation and other harmful practices. The law entered into force in the EU on August 25, 2023, and is aimed at reducing systemic risks and limiting the impact of disinformation. One of the key aspects of the DSA is content moderation. The document states that digital service providers must carefully monitor how their platforms can be used to disseminate manipulative or false content. In addition, each online platform is obliged to provide users with the ability to flag illegal content and report it to the service administration (DSA, 2023).

In China, the Regulations on the Administration of Deep Synthesis of Information Services on the Internet came into force on January 10, 2023. They oblige companies offering tools for creating deepfakes to identify users and provide clear labeling of artificially generated content. This is aimed at preventing misinformation and avoiding confusion among the public (Finlayson-Brown & Ng, 2023).

In Singapore, the Protection Against Online Falsehoods and Manipulation Act (POFMA, 2019 in Singapore Statutes Online, 2020) was passed, giving the government broad powers to require the removal or correction of false content. Although deepfakes are not specifically defined in their law, it covers all types of digital manipulation, including videos or images created by AI, if they are deemed "false" or "misleading". The law is aimed at combating disinformation, but its application has raised concerns among human rights activists. For example, Human Rights Watch claims that POFMA is being used to block critical material and suppress alternative views (Guardian, 2019).

Singapore's experience shows how states can respond quickly to challenges related to diplomatic censorship. However, it is important that such measures do not become a tool to restrict freedom of speech and suppress public debate.

In recent years, South Korea has seen a sharp increase in the number of crimes related to pornographic deepfakes. The police have already registered more than 800 such cases and are actively investigating the activities of Telegram bots that distribute such content. In particular, the impetus for expanding investigations was 88 recorded cases of distributing deepfakes in Telegram. This problem contributed to the adoption of a draft law that criminalizes not only the creation and distribution of such materials, but also their viewing and storage. Previously, such offenses were punishable by up to five years in prison or a fine of up to KRW 50 million (approx. USD 38,461.54[3]) under the Sexual Violence Prevention and Victim Protection Act. However, the new legislation has increased the penalties by raising the maximum prison term to seven years, demonstrating the government's determination to combat such crimes (CNN World, 2024).

The growing number of crimes related to deepfakes and the active work of law enforcement in this area indicate the need for more decisive measures. The introduction of stricter sanctions for such offenses is an important step in the fight against digital manipulation. At the same time, the effectiveness of the resistance depends not only on the severity of the sanctions, but also on the ability of law enforcement agencies to act promptly and effectively in the online environment. Platforms that facilitate the spread of deepfakes play a separate role. Limiting their influence requires a comprehensive approach that combines legal mechanisms and technological tools that can reduce the spread of harmful content.

To better understand the effectiveness of various measures for combating AI-driven disinformation worldwide, it is useful to present them in a comparative Table 2. This

[3] Based on the current exchange rate of USD 1 = KRW 1,300.

table illustrates the combination of legislative, regulatory, and technical approaches employed in different countries and highlights the specific features of their implementation and effectiveness. Such a summary allows for an assessment of which strategies are most successful in reducing the spread of disinformation and enhancing public trust.

Table 2. Comparative overview of legal and technical measures against AI-generated disinformation

Country	Key legal/ regulatory Measures	Technical measures	Targeted content/scope	Observed effectiveness
United States	No Fakes Act (2025; revised draft); State laws such as California AB-730 and Texas SB-751	AI-detection algorithms; platform moderation	Deepfakes affecting elections, media, and personal rights	Moderately effective in limiting election-related disinformation; ongoing challenges with enforcement and platform compliance
European Union	AI Act (2024); Digital Services Act (DSA, 2023)	Mandatory labeling of AI-generated content; watermarks; metadata verification	All AI-generated content, with exceptions for satire, art, and law enforcement	High transparency; improved user awareness; requires further monitoring for enforcement consistency
China	Regulations on Deep Synthesis of Information Services (2023)	User identification; content labeling	AI-generated videos and images	Effective in quickly identifying content sources; concerns regarding freedom of expression
Singapore	Protection Against Online Falsehoods and Manipulation Act (POFMA, 2019)	Government-mandated removal/correction orders	All types of online falsehoods, including AI-manipulated media	Rapid state response; high suppression of disinformation; potential issues with overreach and censorship
South Korea	Criminalization of creation, distribution, and possession of sexually explicit deepfakes	Law enforcement monitoring; Telegram bot investigations	Pornographic deepfakes	Effective in reducing circulation; demonstrates strong deterrence; enforcement relies on active police involvement

Source: Author's own study.

However, the real effectiveness of these strategies is not limited to their technical aspects. In fact, it depends on how each nation defines the problem of deepfakes itself, using specific mechanisms to do so. Two key concepts from communication theory – agenda-setting and framing – help us understand this profound difference.

Agenda-setting explains why the governments of the U.S. and the EU are placing the issue of deepfakes at the highest level, as a threat to democracy and freedom. However, the frame of their fight differs significantly. The United States frames deepfakes as a direct threat to elections and personal rights, which is reflected in its laws aimed

at protecting the political process. The European Union, in turn, views the problem through the broader frame of artificial intelligence transparency, requiring content labeling and verification. This is an approach that empowers users, not just the state.

A completely different situation emerges in China and Singapore. There, the deepfake problem has been put on the agenda as a matter of social control. This frame dictates their approaches: full user identification in China and the rapid suppression of "online falsehoods" in Singapore. These strategies demonstrate high effectiveness in combating content spread, but at the same time, they raise serious concerns about freedom of speech and government overreach.

The situation in South Korea is unique, where deepfakes were placed on the agenda within a narrow frame – that of a moral and criminal offense. The focus on combating sexually explicit deepfakes has allowed the country to achieve significant success in this specific area, demonstrating how a clearly focused frame can ensure decisive and effective countermeasures.

Thus, the analysis of approaches to deepfake regulation goes beyond simple legal or technical assessments. It shows how each country, through the deliberate processes of agenda-setting and framing, not only chooses its method of combat but also defines the balance between control and freedom in the digital age.

In the modern information space, deepfakes and manipulative content are becoming a serious challenge for society. The ability to recognize and counteract false information is not only an individual responsibility, but also a collective task for both citizens and the state, and because of that we have to answer a very current question: how can we effectively combat this problem? Below we offer certain proposals of actions to be applied and developed both at the level of civil society and the state level.

Actions for civil society:

- media literacy education – launching information campaigns to help people learn to recognize deepfakes and other forms of manipulation;
- developing critical thinking – encouraging society to carefully analyze the content disseminated in the media and social networks;
- monitoring and refuting fakes – creating initiatives to identify and expose false information through independent platforms and fact-checking resources.

Actions for the state:

- legal regulation – adoption of regulations establishing liability for the creation and dissemination of deepfakes for manipulation purposes;
- awareness-raising activities – implementation of state initiatives aimed at raising public awareness of the threats of deepfakes and teaching methods of information verification;
- cooperation with technology companies – establishing a partnership with the IT sector to develop transparent standards for labeling content created by AI;

Combating deepfakes requires a comprehensive approach that combines state responsibility with the active engagement of civil society. The government must create

a legal framework and cooperate with technology companies, while citizens must increase their media literacy and develop critical thinking. Only this synergistic approach can ensure effective countermeasures against digital disinformation.

Discussion and conclusion

This study set out to address two key research questions: RQ1: How does AI contribute to disinformation in political PR? and RQ2: What legal and technical solutions exist across different countries to counter AI-driven disinformation?

Regarding RQ1, the analysis demonstrates that AI plays a significant role in generating and disseminating disinformation within political public relations. Artificial intelligence enables the rapid creation of manipulated audiovisual content, including deepfakes, synthetic voices, and algorithmically targeted fake news. Examples such as Barack Obama's 2018 deepfake video, Nancy Pelosi's doctored 2019 video, and the fabricated video of Ukrainian President Volodymyr Zelenskyy illustrate how AI-generated content can distort public perception, manipulate political discourse, and undermine trust in political institutions. These findings align with existing literature, which highlights AI's potential to automate media processes and shape public opinion (Blankenship, 2021; Mehan, 2024), and confirms the ethical risks identified by Pepping et al. (2021) or Lyon and Tora (2023). While previous studies often focus on AI's technical capabilities or societal implications separately, this analysis underscores the intersection of AI technologies with strategic PR practices, demonstrating how AI-driven disinformation actively influences political narratives and public trust.

In addressing RQ2, it is evident that legal and technical responses to AI-generated disinformation vary internationally. The United States has implemented laws such as the No Fakes Act (2025; revised draft) and state-level regulations like California AB-730, which target the misuse of deepfakes in elections and media. The European Union's AI Act (2024) and Digital Services Act (European Commission, 2023) require the labeling of AI-generated content and mandate platform accountability. Asian countries, including China, Singapore, and South Korea, enforce stricter measures, from content identification requirements to criminalization of certain deepfake activities. On the technical front, tools such as Deepware Scanner and Truly Media, alongside watermarking and AI-detection algorithms, are increasingly used to identify and verify manipulated content. These measures, however, are not yet sufficient to fully mitigate the risks, as highlighted by the literature, which notes that most regulatory frameworks remain fragmented and focused on national rather than cross-border strategies.

The findings emphasize that combating AI-driven disinformation requires a multi-layered approach: legal regulations, technical detection tools, proactive PR strategies, and media literacy are all essential. Proactive communication, fact-checking collaborations, and monitoring of media platforms help PR professionals maintain

public trust even in the face of sophisticated AI-generated manipulations. At the societal level, fostering digital resilience skills to critically evaluate, verify, and interpret information is essential for mitigating the impact of disinformation.

In conclusion, AI offers tremendous benefits across sectors, yet its misuse in political PR represents a growing threat to public trust, political stability, and international security. The ultimate challenge is not only technological but also societal: as AI-generated content becomes indistinguishable from reality, collective trust in institutions, media, and factual information is eroded. Moving forward, addressing this challenge requires a combination of robust legal frameworks, advanced technical detection methods, proactive PR practices, and public education, supported by international cooperation. Only through coordinated, comprehensive efforts can societies safeguard information integrity, reinforce democratic processes, and strengthen resilience against AI-driven disinformation.

References

Allyn, B. (2022, March16). *Deepfake video of Zelenskyy could be "tip of the iceberg" in info war, experts warn.* NPR. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

Artificial Intelligence Index Report. (2023). Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

Blankenship, R.J. (2021). *Deep Fakes, Fake News, and Misinformation in Online Teaching and Learning Technologies*. IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-7998-6474-5

California Legislative Information. (2019, April 10). *AB-730 elections: Deceptive audio or visual media. Assembly Bill #730. Chapter 493*. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

CNN World. (2024, September 26). *South Korea to criminalize watching or possessing sexually explicit deepfakes*. https://edition.cnn.com/2024/09/26/asia/south-korea-deepfake-bill-passed-intl-hnk/index.html

Congress.gov. (2024). *NURTURE Originals, Foster Art, and Keep Entertainment Safe (No Fakes) Act*. https://www.congress.gov/congressional-record/congressional-record-index/118th-congress/2nd-session/nurture-originals-foster-art-and-keep-entertainment-safe-no-fakes-act/1920107

Eitren, W. (2024, November 8). *Deep Fakes in the AI Act*. Schjødt. https://schjodt.com/news/deep-fakes-in-the-ai-act

European Commission. (2023). *Digital Services Act*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

European Parliament (2020, September 4). *What is artificial intelligence and how is it used?*. https://www.europarl.europa.eu/topics/en/article/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used

Finlayson-Brown, J., & Ng, S. (2023, February 1). *China brings into force regulations on the administration of deep synthesis of internet technology addressing deepfakes and similar technologies.* A&O Shearman. https://www.aoshearman.com/en/insights/ao-shearman-

on-data/china-brings-into-force-regulations-on-the-administration-of-deep-synthesis-of-internet-technology

Goffman, E. (1986). *Frame Analysis: An Essay on the Organization of Experience*. Northern University Press.

Guardian. (2019, May 9). *Singapore fake news law a "disaster" for freedom of speech, says rights group.* https://www.theguardian.com/world/2019/may/09/singapore-fake-news-law-a-disaster-for-freedom-of-speech-says-rights-group

Horton, J., Sardarizadeh, Sh., & Wendling, M. (2024, August 12). *Trump falsely claims Harris crowd was fake*. BBC. https://www.bbc.com/news/articles/cx2lmm2wwlyo

K2 Communications. (2024). *How Public Relations (PR) Can Combat Fake News?*. https://www.k2communications.in/posts/how-pr-can-combat-fake-news?utm

Konoplianok, V. (2024). *Tekhnolohii politychnykh manipuliatsii za riznykh typiv politychnoho rezhymu*. Natsionalnyi Universytet "Odeska Yurydychna Akademiia".

Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv:1804.08559v1 [cs.SI],* Stanford University, 1–4.

Kuzhman, O.M. (2016). Vstanovlennia poriadku dennoho: mas-media i vlada. *Visnyk Natsionalnoho universytetu "Iurydychna akademiia Ukrainy imeni Yaroslava Mudroho", 3*(30), 274–276.

Legal Information Institute. *10 U.S. Code § 9401 2358 – Research and development projects.* https://www.law.cornell.edu/uscode/text/10/4001

Legal Information Institute. 1*5 U.S. Code § 9401 – Definitions.* https://www.law.cornell.edu/uscode/text/15/9401

LegiScan. (2019). *Bill Text: TX SB751 relating to the creation of a criminal offense for fabricating a deceptive video with intent to influence the outcome of an election*. https://legiscan.com/TX/text/SB751/id/1902830

Lyon, B., & Tora, M. (2023). *Exploring deepfakes: Deploy powerful AI techniques for face replacement and more with this comprehensive guide*. Packt Publishing.

Marchuk, M. (2024). Analiz ta porivniannia metodiv vykorystannia shtuchnoho intelektu v informatsiinii viini ta sposobiv protydii. In *The X International Scientific and Practical Conference "Global achievements and current trends in the development of science"* (pp. 233–236). Bulgaria.

McCarthy, J. (2007). *What Is Artificial Intelligence?* Stanford University.

Mehan, J. (2024). *Artificial Intelligence*. Walter de Gruyter GmbH.

Ministerstwo Cyfryzacji. (2025, January 31). *Pierwsze przepisy Rozporządzenia o sztucznej Inteligencji (AI Act) zaczynają obowiązywać.* https://www.gov.pl/web/cyfryzacja/pierwsze-przepisy-rozporzadzenia-o-sztucznej-inteligencji-ai-act-zaczynaja-obowiazywac

Okhrymovych, V. (2024). Osoblyvosti zastosuvannia dipfeiku v seredovyshchi internet-kultury. Section 7. Modern Media Culture. In *International Scientific Conference* (pp. 47–51). October 3–4, Riga.

Pepping, T., Duivestein, S., & Doorn, M. van, (2021). *Real Fake: Playing with Reality in the Age of AI, deepfakes and the metaverse*. Uitgeverij kleine Uil.

Qiao, F., & Zhou, G., (2024). A deepfakes framework through the lens of framing theory. *OME – An International Journal of Pure Communication Inquiry*.

Rozporiadzhennia Kabinetu Ministriv Ukrainy. (2020, December 2). *Pro skhvalennia Kontseptsii rozvytku shtuchnoho intelektu v Ukraini.* https://zakon.rada.gov.ua/laws/show/1556-2020-p#Text

Silverman, C. (2018, April 17). *How to spot a deepfake like the Barack Obama – Jordan Peele video.* Buzzfeed. https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed

Singapore Statutes Online. (2020). *Protection from Online Falsehoods and Manipulation Act 2019.* https://sso.agc.gov.sg/Act/POFMA2019?ProvIds=P11-#pr1-.

Solis, B. (2023, December 20). *Introducing The GenAI Prism Infographic: A Frame For Collaborating With Generative AI*. https://briansolis.com/2023/12/introducing-the-genai-prism-infographic-a-framework-for-colalborating-with-generative-ai/

Somova, O. (2022, March 9). *Shcho take dipfeik?*. Diia Osvita. https://osvita.diia.gov.ua/news/what-is-a-deepfeak

Soyiba, J., Ullah, A.H., Malik, A.S., & Faye, I. (2019). *Classification of visual and non-visual learners using electroencephalographic alpha and gamma activities.* National Library of Medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC6513874/

The New York Times. (2019, May 24). *Video: Edited Pelosi video vs. the original: A side-by-side*. https://www.nytimes.com/video/us/politics/100000006525055/pelosi-video-doctored.html

The Times of India. (2019, April 12). *FAKE: Viral photo of PM Modi touching Sonia Gandhi's feet.* https://timesofindia.indiatimes.com/times-fact-check/news/fake-viral-photo-of-pm-modi-touching-sonia-gandhis-feet/articleshow/68850648.cms

Trittin-Ulbrich, H., Scherer, A.G., Munro, I., & Whelan, G. (2020). Exploring the dark and unexpected sides of digitalization: Toward a critical agenda. *Organization*, 28(2), 8–25.

Turing, A.M. (1950). *Computing Machinery and Intelligence*. Oxford University Press.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Nechnology Innovation Review*, *9*(11), 39–40.