Coll. Antropol. **49** (2025) 1: 27–35
Original scientific paper
doi:10.5671/ca.49.1.4

AI Hallucinations in Political Contexts: Emerging Challenges to Democratic Trust

Jelena Đuraš Gleđ[1], Domagoj Bebić[1], Nikša Sviličić[2]

[1]Faculty of Political Science, Zagreb, Croatia
[2]Department of Communication, Media and Journalism, University North, Koprivnica, Croatia

ABSTRACT

This paper examines AI-generated hallucinations not only as technical unintended outputs of generative systems but as inherent outcomes of how these models construct meaning and fill informational gaps. In the political sphere, where credibility, accuracy and public trust carry particular weight, such outputs increase the risk of misinformation, heighten polarisation and contribute to growing uncertainty about what is real. Using a qualitative approach that combines a structured review of extant literature with a multi-case analysis of synthetic political content, including short-form deepfake videos and artificial audio disseminated during election periods, the study traces how hallucinations enter and circulate within political communication. The findings show that algorithmic attention systems tend to elevate emotionally charged and personalised material, allowing synthetic content to appear alongside or even above authentic political messages. This blending of sources makes it more difficult for citizens to recognise what is genuine, particularly in fast-paced digital environments. The paper argues that these risks call for a layered response. Regulatory safeguards, clearer provenance and labelling mechanisms, and sustained investment in digital literacy can help limit the democratic harms associated with synthetic media. At the same time, such measures can create space for constructive and transparent uses of AI in political and electoral communication, ensuring that technological innovation does not undermine the foundations of democratic trust.

Key words: *artificial intelligence, AI hallucinations, political communication, deepfakes, electoral process, democracy*

Introduction

The rapid digital transformation of political communication has reshaped how contemporary democracies operate. Over the past decade, artificial intelligence (AI) tools have become intertwined with political marketing, strategic messaging and campaign decision-making. Parties and campaign teams now rely on AI to segment audiences, target voters and automate communication on a scale that was previously impossible[1].

Within this landscape, a new and increasingly consequential challenge has appeared: the rise of AI hallucinations. These arise when generative systems, especially large language models (LLMs), produce text or speech that sounds coherent and authoritative, yet is factually incorrect or entirely invented[2]. In this paper, the term hallucination refers to unintended, system-generated fabrications produced by statistical prediction, whereas deepfakes denote intentionally crafted manipulations created to deceive viewers; distinguishing these two phenomena is essential for analytical and terminological clarity. Hallucinations occur because LLMs do not reason or understand meaning. Instead, they predict the most statistically likely continuation of a sentence based on patterns in their training data[3]. In political contexts, such invented content is not harmless: it can distort how voters interpret events, reinforce misleading narratives and quietly weaken trust in political institutions.

The rapid growth of systems such as ChatGPT, Google DeepMind's Gemini and Meta's LLaMA has accelerated the production and circulation of political content online[4]. As more political engagement unfolds in digital environments organised by recommendation algorithms, the consequences of erroneous AI-generated content become harder to manage[5]. Voters often encounter information selected for its engagement potential rather than its accu-

racy, which allows hallucinated material to spread widely before corrections can catch up[6,7]. Once embedded in the quick tempo of social-media ecosystems, misinformation can shape public perceptions, reinforce pre-existing biases or fuel waves of confusion[8].

This paper investigates how AI hallucinations influence political communication and electoral dynamics. It combines theoretical perspectives with documented examples to examine hallucinations both as computational artefacts and as vulnerabilities within democratic systems. Given these developments, understanding how hallucinations are produced and how they travel across platforms has become essential. This study therefore addresses three research questions. First, how do AI hallucinations and synthetic political content appear across platforms during election periods? Second, which sociotechnical mechanisms enable their visibility, amplification and perceived credibility? Third, what democratic risks arise when such content circulates widely within algorithmically curated information environments?

Literature Review

This chapter synthesises the theoretical, empirical and regulatory scholarship relevant to understanding the intersection of artificial intelligence and political communication. The review covers two interconnected domains: the expanding role of AI within political systems, including its affordances and risks, and the phenomenon of AI hallucinations as a structural property of generative models with growing democratic implications.

By examining existing research across political communication, platform studies, AI ethics and computational linguistics, this chapter provides the conceptual foundation needed to interpret the empirical cases presented later in the article. It addresses how AI systems shape the production and distribution of political content and generate new epistemic vulnerabilities within digital publics.

The role of AI in the political system

Artificial intelligence has become one of the most influential forces in contemporary political communication. Its growing presence is transforming how political actors reach citizens, how decisions are developed, and how public opinion is formed[9,7]. Today, AI systems process large amounts of unstructured data in real time, generate personalised political messages, analyse sentiment, and automate both administrative and communicative tasks[10].

At the same time, the growing influence of AI introduces a number of risks that are difficult to ignore. Targeted political advertising can manipulate information flows, infringe on privacy, and deepen the fragmentation of public discourse. Automated content moderation systems also raise concerns because their decisions are often obscure, shaped by algorithmic bias, and capable of restricting legitimate expression[11]. In addition, AI-assisted decision-making frequently lacks transparency and may function without sufficient human oversight, which weakens the accountability of political actors. The subsections that follow examine the main ways in which AI is applied within political systems. The discussion is structured around five domains.

Personalization of political content

Artificial intelligence enables political actors to design highly targeted messages shaped by users' preferences, behavioural traces and emotional cues. Machine-learning methods make it possible to segment the electorate with increasing precision and adjust communication to ideological profiles, interests and patterns of online engagement[9]. Research on the 2016 U.S. presidential election shows that social-media-based microtargeting played a substantial role in delivering personalised political narratives intended to maximise emotional impact[7,12].

Although such tactics can make campaigns more efficient, they foster the creation of information bubbles and narrow citizens' exposure to diverse viewpoints, weakening pluralism[13]. Most users are not aware that they are being individually targeted by political advertising, which raises concerns about transparency, manipulation and informed consent[14].

Big data analytics

Political actors increasingly rely on AI-assisted analytical tools to interpret large datasets from social networks, digital media and survey research. Advances in natural language processing support trend detection, real-time sentiment analysis and estimates of shifts in voter attitudes[12,15]. These insights enable campaigns to adjust rhetoric and communication strategies more rapidly. During the 2019 European Parliament elections, political actors relied on AI-supported analytical tools to track shifts in issue salience and online debates[9,15]. However, the growing dependence on such systems raises concerns. Models built on historical or incomplete data can unintentionally reinforce existing social biases[16]. Much profiling occurs without citizens being aware of how their data are used, raising questions about transparency, accountability and fairness[13].

Moderation of public discourse

AI tools are now deeply integrated into how platforms manage and control online content. They help identify hate speech, misinformation and harmful communication[17]. Yet the same systems have also played a role in speeding up the spread of misleading political material, particularly during major events[5].

Users often do not understand why certain posts are removed while others remain online, and this lack of clarity fuels doubts about fairness and accuracy. In some cases, political messages are taken down while near-identical content stays visible, producing inconsistency and frustration[11,18]. Content moderation supported by AI therefore

occupies an ambivalent position. While it contributes to maintaining safer online spaces, it also provokes debates about fairness, accountability and the scope of legitimate democratic expression.

Algorithmic decision-making

Artificial intelligence is integrated into many routine tasks in political and administrative work. Campaign teams use automated systems to organise outreach, schedule posts, coordinate volunteers and respond to inquiries. Chatbots and automated email tools help handle repetitive communication[11]. One illustrative example is Estonia's national AI assistant, Kratt, designed to help people navigate public services through a single digital entry point[19].

However, automating political communication creates complications. Citizens tend to react negatively when responses appear formulaic or detached from context. Automated systems may deliver incomplete or misleading information, raising questions about responsibility and oversight[20].

Public institutions increasingly adopt AI to support routine administrative work. These tools can sort applications, support decisions about resource distribution and handle routine processing. Yet many AI models operate as black boxes, with outcomes that may be uneven or discriminatory, gradually eroding public trust[16,21]. A newer development is the appearance of synthetic political actors: AI-generated personas, automated commentators and artificial profiles that engage in political discussions without disclosure[22]. These actors blur the boundary between authentic and artificial political speech.

AI hallucinations: mechanisms, typology and electoral relevance

AI hallucinations constitute one of the most significant limitations of generative models in political communication. The term describes the capacity of large language models and multimodal systems to produce outputs that are linguistically coherent yet factually incorrect[2,4]. Hallucinations are structural artefacts arising from the statistical principles that govern generative AI, which makes their study essential.

Hallucinations primarily emerge because LLMs do not possess semantic understanding but generate responses by estimating probable continuations[3,23]. When contextual cues are incomplete, ambiguous or time-sensitive, the model fills informational gaps through pattern completion[2]. Another mechanism is overgeneralisation, especially in political prompts dominated by scandals, corruption narratives or geopolitical tensions[24]. Retrieval failure and contextual drift occur when prompts require up-to-date or specific factual knowledge[25,26].

Conversational optimisation increases risks: generative systems maximise coherence rather than truthfulness, continuing fabricated elements to maintain narrative flow[27,28]. In multimodal systems, hallucinations appear as visual or auditory artefacts[29]. Repeated exposure reduces detection ability, increasing perceived authenticity[30]. Even low-quality synthetic media can generate political consequences when circulating in high-velocity environments[31].

Hallucinations in political contexts appear in several forms: factual hallucinations, attributive hallucinations, logical hallucinations, predictive hallucinations[32], multimodal hallucinations[33] and bias-induced hallucinations[34]. These forms illustrate that hallucinations are sociotechnical distortions shaped by platform design, training data and user interaction. Fabricated or distorted content can affect how voters understand political actors, policies or polling expectations, particularly before elections[35,36].

Short-form platforms heighten these risks. Fast-scrolling environments reduce authenticity evaluation, while emotionally charged content gains disproportionate visibility[37,38]. Recommendation systems privilege emotionally arousing or provocative content[6,39]. A long-term consequence is the erosion of epistemic trust. As synthetic media become more realistic, citizens may become uncertain about what to believe, enabling political actors to dismiss legitimate evidence as fake[5,7]. Recent incidents, such as the Biden robocall in 2024 and the synthetic audio in the 2023 Slovak elections, demonstrate the electoral relevance of hallucinated content[33].

AI hallucinations are sociotechnical vulnerabilities embedded within digital platforms. Their migration into audiovisual formats, combined with algorithmic amplification and declining public trust, increases their relevance for electoral processes and democratic legitimacy. These dynamics relate to the second research question on visibility, plausibility and impact. Taken together, the literature demonstrates that AI reshapes political communication through intentional technological affordances and unintended generative distortions. These dual dynamics underscore the need to analyse hallucinations as sociotechnical artefacts within digitally mediated political environments.

Methodology

This study employs a qualitative, multi-layered methodological approach designed to analyse how AI hallucinations shape political communication, electoral processes and platform-based information flows. The research integrates three methodological components: a structured literature review, a comparative multi-case study analysis and a conceptual analytical framework grounded in political communication theory, platform studies and AI ethics.

Structured literature review

A structured literature review was conducted to synthesise academic, policy and technical knowledge relevant to AI hallucinations, political misinformation and electoral communication. The review included peer-reviewed journal articles, scientific reports and regulatory documents published between 2018 and 2024. This in-

cluded literature on hallucinations in large language models and the ways in which these systems produce confident but inaccurate outputs[3,2], studies analysing how political communication unfolds on digital platforms[40,7,13], and documents related to the EU AI Act and the Digital Services Act[41,42]. The literature set also included research on deepfakes, audiovisual manipulation and the psychological factors shaping how people deal with misinformation[32,31].

Inclusion criteria required relevance to AI-generated misinformation, electoral contexts or sociotechnical risk. The review provided the theoretical foundation for understanding hallucinations not only as computational artefacts but as communication phenomena embedded in platform architectures. The analysis focused on three thematic clusters addressing AI hallucinations and generative model behaviour, including studies examining the origins, typologies and mechanisms of hallucinations in large language models and multimodal systems[3,2]; misinformation, political persuasion and digital platform dynamics, drawing on research exploring how digital infrastructures shape misinformation diffusion, political behaviour and voter decision-making[32,6,7]; and AI governance, transparency and ethical regulation, including policy frameworks and scholarly analyses on the EU AI Act, the Digital Services Act and normative principles such as explainability, accountability and provenance[41,42,21].

Multi-case study approach

A comparative multi-case study approach was employed[43]. Cases were selected according to four criteria: political salience, presence of AI-generated or hallucinated synthetic content, documented public impact and media format diversity. The selection process involved compiling a pool of publicly documented incidents involving synthetic political content and including only cases with verifiable traces of AI-generated manipulation and reconstructable cross-platform circulation. Each case was analysed through a uniform three-layer protocol including technical examination of audio-visual artefacts, tracing platform-specific dissemination patterns and interpreting discursive effects in relation to political narratives and media-system characteristics.

Three cases were selected to capture variation in format, political context and type of AI-generated distortion. The analysis first focuses on technical traces that typically accompany AI-generated media; the second part examines platform-level dynamics, in particular how TikTok's recommendation system can push emotionally charged clips to wider audiences[38]; and the third part situates the cases within wider theoretical discussions, including epistemic vulnerability[6], algorithmic virality and affect in political communication. These cases differ in media format, political timing and democratic relevance. Together, they offer a diverse empirical foundation suitable for examining how AI-generated political content functions across different platform environments.

Conceptual analytical framework

The analysis is guided by an interdisciplinary conceptual framework integrating political communication theory, insights from algorithmic attention economies and principles from AI ethics and governance. Political communication theory provides concepts such as agenda-setting, framing, mediatization and affective polarisation[44,45]. Platform studies offer insights into how recommendation systems reward novelty, emotional intensity and humour over factual accuracy[38,11], shaping the visibility of synthetic political content. Normative principles from AI ethics, transparency, accountability, explainability and provenance tracking[21] inform the evaluation of risks and safeguards relevant to understanding the democratic implications of AI-generated media.

Limitations

The study does not include quantitative experiments or survey data capable of measuring behavioural effects or individual susceptibility to hallucinated content. The case study approach relies on publicly documented incidents, which may omit covert interference operations or unseen platform moderation. Additionally, the speed of generative AI development may outpace some conceptual frameworks included in the review. Despite these limitations, the triangulated approach offers strong analytical depth and captures the sociotechnical mechanisms through which hallucinations gain visibility and political relevance.

Results

The three cases examined in this study offer a clearer picture of how synthetic political material appears, circulates and gains relevance within different digital environments. Although the contexts differ considerably, the same mechanisms seem to surface repeatedly.

Comparative overview of the cases

The cases from Croatia, Slovakia and the United States illustrate three distinct uses of AI-generated political content. One revolves around humorous reinterpretation, another around strategically timed manipulation, and the third around an attempt to influence voter behaviour directly.

Technical indicators

Despite the differences between video, audio and robocalls, all three examples contained identifiable artefacts. The specific artefacts varied, but none prevented wide circulation. Imperfections did not reduce the visibility of the material, and voice-based manipulations were generally more convincing than visual ones.

Platform dynamics

Each platform created a different pathway through which synthetic political content circulated. TikTok am-

TABLE 1

APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN THE POLITICAL SYSTEM: BENEFITS, RISKS, AND SAFEGUARDS

AI Application	Benefits	Risks / Drawbacks	Possible Safeguards
Personalization of political content	Message relevance; effective campaigns	Microtargeting; polarisation; information bubbles	Transparency; advertising restrictions; civic education
Big data analytics	Rapid sentiment and trend analysis; predictive insights	Bias reproduction; lack of explainability	Algorithmic accountability; human oversight; data audits
Content moderation	Reduced hate speech; improved online safety	Censorship; bias; opaque decisions	Clear regulatory frameworks; human oversight
Automation of political communication	Efficiency; scalability; cost reduction	Loss of authenticity; manipulation; unclear accountability	Labeling AI-generated content; ethical standards
AI-based decisions in administration	Faster processing; reduced administrative burden	Opaque and biased decisions; reduced legal protection	Human-in-the-loop; right to explanation and appeal
AI hallucinations in political discourse	Rapid content generation when supervised	Disinformation; loss of trust in institutions	Verification systems; uncertainty labels; AI watermarking

plified clips based on engagement signals, enabling even lightly manipulated videos to reach large audiences. Telegram facilitated rapid spread within ideologically aligned or closed communities, giving synthetic audio high initial momentum. Robocalls bypassed platform infrastructures entirely, drawing their perceived legitimacy from the communication channel itself. These platform-specific dynamics shaped exposure and influence more strongly than the technical characteristics of the synthetic content.

Discursive effects

The three cases generated distinct discursive effects once synthetic content entered public circulation. Although they differed in tone, intent and medium, several common patterns emerged. In Croatia, humorous reinterpretations encouraged audiences to treat political communication as entertainment, blurring the boundary between satire and reality. This contributed to a gradual normalisation of synthetic portrayals and a more playful, yet increasingly cynical, perception of political actors.

In Slovakia, synthetic audio interacted with existing narratives about corruption and political distrust. The timing of the material released during media silence amplified its discursive force, allowing the fabricated message to shape public conversation before corrective information could gain traction.

In the United States, the voice-cloned robocall drew its power from the authority cues embedded in the president's voice. Synthetic content operated not as satire or rumour, but as a direct intervention aimed at influencing voter behaviour, raising concerns about the vulnerability of democratic participation to audio-based manipulation. Across all three contexts, synthetic content shaped interpretive frames, emotional responses and trust toward political institutions.

Narrative case descriptions

Croatia

In the months before the 2024 parliamentary elections, a series of TikTok videos appeared in which footage of Prime Minister Andrej Plenković was matched with humorous, synthetic audio and lip-sync overlays. Viewers could usually spot small inconsistencies, slight delays in lip movement, and subtle mismatches between facial expressions and audio, but these details did not hinder circulation. Clips quickly surfaced in feeds that mixed political content with comedy and general entertainment, and later migrat-

TABLE 2

OVERVIEW OF THE THREE CASES

Case	Country	Format	Intent/Framing	Timing	Observed Effect
1	Croatia	TikTok short-form video	Humorous / satirical	Pre-election period, 2024	Normalisation of altered portrayals; overlap of satire and political content
2	Slovakia	AI-generated audio	Fabricated vote-buying narrative	48 hours before 2023 election	Confusion during media silence; rapid spread
3	U.S.	Voice-cloned robocall	Voter-suppression attempt	Day before 2024 NH primary	Attempted behavioural influence via impersonation

ed to Instagram Reels, Facebook pages and private group chats. Although clearly humorous, users repeatedly debated whether the original statement had in fact occurred, illustrating how playful distortions can momentarily destabilise memory of authentic political communication.

Slovakia

Two days before the 2023 parliamentary election, an audio recording appeared in several large Telegram channels. The voices sounded like opposition politicians allegedly discussing vote-buying, although irregular pacing and tonal inconsistencies were detectable. The recording spread quickly because the media silence period prevented immediate rebuttal. It soon migrated to TikTok and Facebook, where commentary pages and private groups reshared it. Shorter excerpts stripped of context circulated independently, intensifying confusion even among users who had not heard the full audio.

United States

In New Hampshire, on the eve of the 2024 Democratic primary, registered voters received automated phone calls in which a voice resembling President Joe Biden urged them not to vote. The imitation matched Biden's vocal timbre, tempo and phrasing unusually closely. Because robocalls bypass social-media infrastructures, the message avoided moderation entirely. Local officials responded within hours, yet the incident highlighted how direct-to-voter channels can be exploited at moments when time for correction is limited.

Cross-case synthesis

Across the three examples, several recurring empirical patterns emerged. Imperfection rarely limited spread, platform design shaped visibility, authenticity cues were weakened by context collapse, timing mattered, authority signals increased persuasive force, and content rarely stayed on one platform. Taken together, the results show how synthetic political content gains relevance not only because of how it is produced, but because of how platforms distribute and frame it once it is released. The three cases demonstrate that the combination of technical artefacts, platform architecture and audience expectations produces the conditions under which synthetic content can acquire political significance.

Discussion

This article set out to examine how AI hallucinations and synthetic political content circulate across contemporary digital ecosystems, and how their platform-driven visibility shapes political perception and democratic processes. Building on the three empirical cases and the three-layer analytic model introduced in the methodology, this section connects the results with the theoretical and regulatory debates presented in the literature review. The central argument that emerges is that hallucinations are not isolated technical glitches, but sociotechnical distortions that operate inside attention-driven, hybrid and emotionally charged media environments.

AI hallucinations as sociotechnical phenomena

At first sight, AI hallucinations may look like a purely technical problem: language models and multimodal systems sometimes make mistakes. Generative models produce output through statistical prediction and pattern completion rather than semantic understanding; when prompts are ambiguous, time-sensitive or under-specified, models tend to fill gaps with fabricated but fluent content[3,2]. Overgeneralisation from skewed training data, retrieval failures and temporal drift further increase the

TABLE 3

COMPARATIVE EXAMINATION OF THE POLITICAL IMPACT OF AI-DRIVEN MANIPULATIONS

Case	Country	Format / Medium	Type of Synthetic Distortion	Context & Mechanism	Observed Democratic Impact
Case 1	Croatia	TikTok short-form video	Lip-sync overlays, synthetic voice cues, humorous distortions	Algorithmic amplification of emotionally engaging satire; co-existence of synthetic and authentic content on users' feeds	Normalisation of synthetic portrayals; erosion of certainty about political statements; blurring boundaries between satire and political information
Case 2	Slovakia (2023 elections)	AI-generated audio clip	Fabricated conversation (vote-buying hallucination)	Released 48h before elections; rapid spread via Telegram → TikTok → Facebook	Real-time electoral interference; public confusion; weakened trust in the electoral process
Case 3	United States (2024 NH primary)	Robocall voice-clone	Biden voice impersonation; synthetic authority cue	Direct-to-voter distribution exploiting trust in official robocall format	Potential turnout suppression; impersonation of political authority; behavioural manipulation

likelihood of invented statements, misattributed quotes or entirely fictional events[24–26].

The cases analysed here show that these mechanisms do not operate in a vacuum. In multimodal systems, hallucinations take on visual or sonic form–slightly off lip-sync, irregular facial micro-expressions, odd prosody or subtle artefacts typical of consumer-level deepfake tools[32]. Imperfections did not prevent the content from travelling widely or shaping interpretation. In fast-scrolling environments, where users rarely pause to verify authenticity, minor artefacts are easy to miss or simply treated as part of the platform's aesthetic[30].

Across all three cases, hallucinations appear as socio-technical phenomena. They originate in model architecture and training data, but their political relevance depends on how they are embedded in platform design, user behaviour and local political context. Environments that reward emotional, humorous or sensational content make hallucinated or synthetic material more likely to find an audience and accumulate political meaning[45,38].

The Croatian case: normalising synthetic portrayals

The Croatian example illustrates how ostensibly harmless, humorous content can gradually acquire political weight. TikTok videos that remix or lip-sync footage of Prime Minister Andrej Plenković are framed as jokes, yet they frequently appear alongside genuine political clips. Users searching for authentic speeches routinely encountered a mixture of verified and synthetic material, a direct consequence of a recommendation system that privileges engagement signals over informational accuracy[38,17].

In a small and centralised media system such as Croatia's, this co-presence of authentic and synthetic portrayals blurs interpretive boundaries. Over time, it becomes harder to recall what was actually said in a specific speech and what belongs to a meme or remix. Repeated exposure contributes to habituation to deepfakes and synthetic media: the more often people encounter such content, the less cognitively vigilant they become[30,40]. Humour plays a central role, lowering analytical guard and encouraging heuristic processing[46].

The Croatian case unfolds in a context marked by relatively low trust in institutions and traditional media[47]. Synthetic portrayals that depict politicians as absurd or contradictory can easily resonate with pre-existing cynicism[48]. Content does not stay confined to a single platform; clips posted on TikTok are quickly reposted to Instagram Reels, Facebook and messaging apps[49]. As they migrate, they lose contextual cues that originally signalled satire, increasing interpretive ambiguity. Even non-malicious distortions can contribute to a slow normalisation of synthetic portrayals in everyday political talk.

TABLE 4

TECHNICAL AND PLATFORM-LEVEL CHARACTERISTICS ACROSS CASES

Dimension	Croatia	Slovakia	U.S.
		Technical traces	
Lip-sync mismatch	✓	–	–
Facial-expression inconsistencies	✓	–	–
Irregular vocal timbre	✓	✓	minimal
Prosodic irregularities	–	✓	✓
Consumer-level generation artefacts	✓	✓	–
High-fidelity cloning	–	–	✓
		Platform dynamics	
TikTok recommendation loops	Very strong	Moderate	–
Telegram group virality	–	Strong	–
Cross-platform spread	Frequent	Frequent	–
Direct-to-user delivery	–	–	Strong
Mismatch between user intent and results	Strong	Moderate	–
		Discursive outcomes	
Blurring satire with real politics	✓✓✓	–	–
Reinforcing corruption narratives	–	✓✓	–
Undermining trust in institutions	✓	✓	✓
Exploiting authority cues	–	✓	✓✓✓
Direct electoral interference	–	✓	✓✓✓

Strategic manipulation in Slovakia and the United States

The Slovak and U.S. cases represent more strategic uses of synthetic media during electoral processes. In Slovakia, an AI-generated audio recording purporting to capture opposition leaders planning vote-buying appeared during the media-silence period. Recognisable voices and a storyline echoing familiar corruption narratives lent the clip plausibility[33]. Telegram's architecture facilitated rapid closed-group virality before the audio spilled over to TikTok and Facebook[33,6]. Media silence limited journalistic and institutional responses, allowing misleading content to shape interpretation before corrections appeared[6].

The New Hampshire robocall episode reveals another manipulation vector. A high-fidelity voice clone of President Joe Biden instructed voters not to participate in the primary. Dissemination relied on robocall systems historically coded as official political communication. This "synthetic authority" was central[50]. Both cases show synthetic media intervening directly in electoral behaviour, exploiting institutional rules and communication channels.

Cross-case synthesis: converging mechanisms

Read together, the three cases reveal recurring mechanisms. Technical imperfection is not a barrier to influence. Platform and channel logics outweigh content characteristics. Context collapse blurs interpretive boundaries[52,53]. Temporal proximity to elections heightens vulnerability[6,33]. Cross-platform migration extends the life cycle of synthetic content[49]. These mechanisms point to a broader shift: AI hallucinations introduce not only discrete falsehoods but also a more diffuse form of epistemic instability[34].

Democratic implications

The democratic stakes are substantial. The first risk concerns epistemic trust. As synthetic videos, voice-cloned messages and hallucinated statements circulate more frequently, citizens encounter more cases where what appears authentic later proves fabricated[32]. This creates fertile ground for the "liar's dividend": public awareness of deepfakes makes it easier for political actors to dismiss genuine evidence as fake[31].

A second implication concerns emotionalisation. Algorithmically curated environments reward strong affective responses[7,8]. Synthetic media integrate seamlessly because they can be tailored for maximum surprise or provocation. The Croatian case shows humour-driven effects; the Slovak and U.S. cases demonstrate fear or confusion triggered through fabricated audio[33,54].

A third risk lies in new vectors for electoral interference. Voice-cloned messages exploit trust historically attached to familiar voices and institutional formats[53]. When introduced at moments of high salience, there is little time to verify authenticity[6,33].

Existing governance frameworks face limitations. Regulatory initiatives such as the EU's AI Act introduce classifications and transparency obligations, but implementation lags behind innovation[21]. Platform-level measures remain fragmented[24]. Digital and AI literacy emerges as a crucial safeguard[8].

Contribution to theory and research

The findings contribute in several ways. They reconceptualise hallucinations as communicative phenomena shaped by platform incentives, audience expectations and cultures of trust[34,45]. They extend hybrid media systems theory to AI-mediated audiovisual manipulation[45]. The Slovak and U.S. cases identify synthetic authority as an emergent mechanism of political influence[33,54]. Cross-platform trajectories invite further research combining computational tracking with qualitative fieldwork[40,49].

In sum, AI hallucinations and synthetic political media should be understood not only as technical challenges, but as catalysts of deeper transformations in how democratic societies produce, circulate and evaluate political information.

Conclusion

This paper examined how AI hallucinations and synthetic political media shape contemporary political communication, focusing on their manifestations across digital platforms, the sociotechnical mechanisms that amplify their visibility, and the democratic risks that arise from their circulation. By following three different episodes, a cycle of humorous TikTok edits in Croatia, a strategically released audio deepfake just before the Slovak elections, and a voice-cloned robocall targeting voters in the United States, the study traced how synthetic political content gains traction and why it matters for democratic life.

Taken together, these cases show that hallucinated or manipulated material does not circulate in a vacuum. It enters digital spaces that are already shaped by recommendation systems, emotional engagement cues and habitual modes of media consumption. These structures determine not only what users see, but also what they assume to be credible. This helps explain why even crude manipulations can become politically meaningful: the speed of circulation, the blending of entertainment and information, and the lack of verification cues often override the technical flaws of the content itself.

The first research question asked how AI-generated political content appears across platforms. The findings show that its forms are diverse. Some pieces emerge from generative models as spontaneous errors; others are created deliberately to mislead. Their presentation ranges from playful and ironic to overtly strategic, and they frequently coexist with authentic material in ways that blur interpretative boundaries.

The second question concerned the sociotechnical mechanisms that give such content visibility. Across the

cases, platform architecture proved more decisive than the content's intrinsic qualities. TikTok elevated humorous distortions because they fit its engagement logic; Telegram allowed rapid spread within closed communities; robocall systems gave synthetic audio a veneer of institutional authority. What matters is not only the manipulation itself but also the infrastructure through which it travels.

The third question asked what democratic risks follow from these dynamics. The evidence points to several. In Slovakia and the United States, synthetic audio exploited moments when voters were most vulnerable to confusion. In Croatia, repeated exposure to manipulated portrayals fed into broader scepticism toward political communication. These cases underscore the possibility that the mere presence of synthetic media, even when recognised as such, can weaken shared factual ground and make it easier for political actors to dismiss unwelcome evidence.

From a theoretical perspective, the study suggests that hallucinations should be treated not simply as faults within AI systems but as distortions that gain meaning only when situated within specific platform and cultural contexts. Algorithmic curation, hybrid media systems and existing public attitudes all shape the political consequences of synthetic content. This reframing shifts the discussion from a purely technical problem toward a sociotechnical one.

The study also has limits. It focuses on documented incidents, meaning that covert or less-visible manipulations remain outside the scope. It also cannot determine how individual users interpret or act on synthetic content, which would require experimental or survey work. Future research could therefore combine case studies with behavioural evidence and computational analyses of platform flows to build a fuller picture.

Despite these constraints, the findings point to a clear conclusion: as generative AI becomes woven into everyday communication, democratic safeguards must evolve with equal speed. Technical provenance tools, transparent platform governance and sustained public education are all part of the response. Above all, democracies will need strategies that help citizens maintain confidence in what they see and hear. Without such efforts, the growing presence of synthetic political content risks undermining the informational foundations on which democratic decision-making depends.

REFERENCES

1. ANSTEAD N, The Political Quarterly, 92 (2021) 282. doi:10.1111/1467-923X.12934. — 2. JI Z, LEE N, FRIESKE R, YU T, SU D, XU Y, SUN Y, ACM Comput Surv, 55 (2023) 1. doi:10.1145/3571730. — 3. BENDER EM, GEBRU T, MCMILLAN-MAJOR A, SHMITCHELL S, Proc ACM Conf Fairness Accountability Transparency, 2021, 610. doi:10.1145/3442188.3445922. — 4. WEIDINGER L, MELLOR J, RAUH M, GRIFFIN C, UESATO J, GLAESE A, GABRIEL I, Proc ACM Conf Fairness Accountability Transparency, 2022. doi:10.1145/3531146.3533088. — 5. ZHAO W, MENG F, LIU X, YU H, WANG Z, WANG H, ACM Trans Intell Syst Technol, 14 (2023) 1. doi:10.1145/3554739. — 6. VERGEER M, Inf Commun Soc, 23 (2020) 1500. doi:10.1080/1369118X.2020.1772862. — 7. WEST DM, ALLEN JR, CHEN A, Artificial Intelligence and Democracy: Risks and Promises (Brookings Institution, Washington, 2022). — 8. GORWA R, BINNS R, KATZENBACH C, Big Data Soc, 7 (2020) 1. doi:10.1177/2053951719897945. — 9. BRENNEN JS, SIMON F, HOWARD PN, NIELSEN RK, Types, Sources, and Claims of COVID-19 Misinformation (Reuters Institute, Oxford, 2020). — 10. LEWANDOWSKY S, COOK J, ECKER UKH, ALBARRACÍN D, AMAZEEN MA, KENDEOU P, VRAGA EK, PLOS Clim, 1 (2022) e0000002. doi:10.1371/journal.pclm.0000002. — 11. ZUBOFF S, The Age of Surveillance Capitalism (PublicAffairs, New York, 2019). — 12. GUESS A, NYHAN B, REIFLER J, Nat Hum Behav, 4 (2020) 472. doi:10.1038/s41562-020-0833-x. — 13. KREISS D, MCGREGOR SC, Political Commun, 35 (2018) 155. doi:10.1080/10584609.2017.1364811. — 14. BALDWIN-PHILIPPI J, Internet Policy Rev, 8 (2019) 4. doi:10.14763/2019.4.1437. — 15. TUFEKCI Z, First Monday, 19 (2014) 7. — 16. ZUIDERVEEN BORGESIUS FJ, TRILLING D, MÖLLER J, BODÓ B, DE VREESE CH, HELBERGER N, Internet Policy Rev, 9 (2020) 1. doi:10.14763/2020.1.1503. — 17. ANANNY M, CRAWFORD K, New Media Soc, 20 (2018) 973. doi:10.1177/1461444816676645. — 18. LAZER DMJ, BAUM MA, BENKLER Y, BERINSKY AJ, GREENHILL KM, METZGER MJ, ZITTRAIN JL, Science, 359 (2018) 1094. doi:10.1126/science.aao2998. — 19. LEPRI B, OLIVER N, LETOUZÉ E, PENTLAND A, VINCK P, Philos Technol, 34 (2021) 567. — 20. GILLESPIE T, Media Cult Soc, 41 (2019) 175. doi:10.1177/0163443718798902. — 21. HWANG T, TADDEO M, FLORIDI L, AI Soc, 38 (2023) 145. doi:10.1007/s00146-022-01483-0. — 22. VINNAL K, ROSS J, Gov Inf Q, 38 (2021) 101693. doi:10.1016/j.giq.2021.101693. — 23. JUNG K, PARK S, Gov Inf Q, 38 (2021) 101608. doi:10.1016/j.giq.2021.101608. — 24. FLORIDI L, COWLS J, BELTRAMETTI M, et al., Minds Mach, 28 (2018) 689. doi:10.1007/s11023-018-9482-5. — 25. HELBERGER N, PIERSON J, POELL T, Inf Commun Soc, 23 (2020) 174. doi:10.1080/1369118X.2019.1583918. — 26. LIN S, HILTON J, EVANS O, Proc Annu Meet Assoc Comput Linguist, 2022, 7039. doi:10.18653/v1/2022.acl-long.229. — 27. MAYNEZ J, NARAYAN S, BOHNET B, MCDONALD R, Proc 58th Annu Meet Assoc Comput Linguist, 2020, 1906. — 28. DZIRI N, KAMALLOO E, YAVUZ S, ZAIANE O, Findings Assoc Comput Linguist, 2022, 243. doi:10.18653/v1/2022.findings-acl.131. — 29. LEWIS P, PEREZ E, PIKTUS A, PETRONI F, KARPUKHIN V, GOYAL N, RIEDEL S, Adv Neural Inf Process Syst, 33 (2020) 9459. — 30. SHUSTER K, JU D, ROLNICK D, SZLAM A, WESTON J, arXiv, 2021, 2109.05014. — 31. PÉREZ J, KIELA D, WESTON J, Trans Assoc Comput Linguist, 10 (2022) 716. doi:10.1162/tacl_a_00492. — 32. KORSHUNOV P, MARCEL S, arXiv, 2018, 1812.08685. — 33. DE LEYN T, FRISSEN T, PUTZEYS T, DE WOLF R, VAN DEN BULCK J, Comput Hum Behav, 139 (2023) 107524. doi:10.1016/j.chb.2022.107524. — 34. CHESNEY R, CITRON DK, Calif Law Rev, 107 (2019) 1753. — 35. PENNYCOOK G, RAND DG, Trends Cogn Sci, 25 (2021) 388. doi:10.1016/j.tics.2021.02.007. — 36. SCHICK N, J Democr, 35 (2024) 112. — 37. CHADWICK A, The Hybrid Media System: Politics and Power (Oxford University Press, Oxford, 2017). — 38. COTTER K, SMITH A, Social Media Soc, 8 (2022) 1. doi:10.1177/20563051221122812. — 39. ALLCOTT H, GENTZKOW M, J Econ Perspect, 31 (2017) 211. doi:10.1257/jep.31.2.211. — 40. TUCKER JA, GUESS A, BARBERA P, VACCARI C, SIEGEL A, SANOVICH S, NYHAN B, Soc Media Polit Polarization Polit Disinform, 1 (2018) 1. — 41. HIGHFIELD T, Social Media Entertainment and Politics (Polity, Cambridge, 2022). — 42. MIHELJ S, KÜBLER M, Journal Stud, 21 (2020) 923. doi:10.1080/1461670X.2020.1745669. — 43. VACCARI C, CHADWICK A, Social Media Soc, 6 (2020) 1. doi:10.1177/2056305120903408. — 44. FLORIDI L, COWLS J, BELTRAMETTI M, et al., Nat Mach Intell, 5 (2023) 176. doi:10.1038/s42256-023-00631-w. — 45. HELBERGER N, BASTIAN M, MICKLITZ HW, SAX M, Internet Policy Rev, 11 (2022) 1. doi:10.14763/2022.1.1611. — 46. YIN RK, Case Study Research and Applications: Design and Methods (SAGE, Thousand Oaks, 2018). — 47. MCCOMBS M, Journal Stud, 6 (2005) 543. doi:10.1080/14616700500250438. — 48. PERUŠKO Z, VOZAB D, Javnost, 28 (2021) 1. doi:10.1080/13183222.2020.1859548. — 49. LAMARRE HL, WALTHER WO, Commun Res,

40 (2013) 303. doi:10.1177/0093650211416648. — 50. EUROPEAN COMMISSION, Standard Eurobarometer 100: Public Opinion in the European Union (Publications Office of the European Union, Luxembourg, 2023). — 51. ŠTULHOFER A, BAĆAK V, ŠEVCÍKOVÁ A, Polit Psychol, 41 (2020) 813. doi:10.1111/pops.12655. — 52. VOZAB D, PAVLOVIĆ D, SUŠAC V, Medijske studije, 13 (2022) 1. — 53. PAWELEC B, Media Cult Soc, 44 (2022) 1711. — 54. MONTAG C, DUKE É, MARKOWETZ A, Addict Behav Rep, 14 (2021) 100371. doi:10.1016/j.abrep.2021.100371.

N. Sviličić

University North, Ul. 104. brigade 3, 42000 Varaždin, Croatia
e-mail: niksa.svilicic@proactiva.hr

HALUCINACIJE UMJETNE INTELIGENCIJE U POLITIČKIM KONTEKSTIMA: NOVI IZAZOVI ZA DEMOKRATSKO POVJERENJE

S A Ž E T A K

Ovaj rad razmatra halucinacije generirane umjetnom inteligencijom ne samo kao tehničke, nenamjerne izlaze generativnih sustava, nego kao inherentne posljedice načina na koji ovi modeli konstruiraju značenje i popunjavaju informacijske praznine. U političkoj sferi, u kojoj vjerodostojnost, točnost i povjerenje javnosti imaju posebnu težinu, takvi izlazi stvaraju rizike koji mogu destabilizirati percepciju stvarnosti, potaknuti polarizaciju i oslabiti povjerenje u institucije. Koristeći kvalitativni pristup koji kombinira strukturirani pregled literature i analizu više slučajeva sintetičkog političkog sadržaja, uključujući kratke videozapise nastale dubinskom sintetičkom obradom i umjetno generiran zvuk distribuiran tijekom izbornih razdoblja, istraživanje prati načine na koje halucinacije ulaze u političku komunikaciju i kako unutar nje cirkuliraju. Nalazi pokazuju da digitalna okruženja usmjerena na privlačenje pozornosti daju prednost emocionalno nabijenim i personaliziranim sadržajima, zbog čega se sintetički materijali pojavljuju uz ili iznad autentičnih političkih poruka. Takvo preklapanje izvora otežava građanima razlikovanje stvarnog od umjetno proizvedenog, osobito u brzim i informacijski zasićenim medijskim prostorima. Analiza sugerira da ovakvi rizici zahtijevaju višeslojni odgovor: regulatorne mehanizme, jasne oznake izvornosti i podrijetla sadržaja te sustavno ulaganje u digitalnu i medijsku pismenost. Takav pristup može ograničiti demokratsku štetu povezanu sa sintetičkim medijima i istovremeno očuvati temeljno povjerenje građana u političku komunikaciju.