CEE Digital Democracy Watch

The Road to the Ballot.

AI Governance and the Integrity of Democratic Choice

The Road to the Ballot.
AI Governance and the Integrity of Democratic Choice

Authors: Konrad Kiljan, Jakub Szymik

External contributors: Marta Bieńkiewicz , Anthony DeMattee, Agata Hajduk-Smak, Chinasa T. Okolo, Samuel Stockwell, Aleksandra Wójtowicz

Editing: Miles Maftean
Graphic design: NIKIFOR.studio

Made possible by the support of Civitates.

Civitates
The European Democracy Fund

Number ISBN 978-83-978606-1-2

ISBN 978-83-978606-1-2
9 788397 860612

CEE Digital Democracy Watch
Fundacja Obserwatorium Demokracji Cyfrowej
Mokotowska 43/104, 00-551 Warszawa
KRS: 0001090110 REG: 114284692189-05
www.ceeddw.org

Table of Contents

Introduction

In 2024, a group of leading AI companies met in Munich to sign a declaration on election integrity, effectively acknowledging that the rapid development of their tools could no longer be treated as politically neutral. The declaration reflected a growing awareness that AI systems were already moving into the core of electoral communication, campaigning, and information access. Yet the commitments remained voluntary, and they sat alongside a broader but still fragmented regulatory landscape, in which some jurisdictions were beginning to develop risk-based rules for high-risk systems, transparency, and accountability, while others were progressing more slowly.

One year later, it was already clear that voluntary pledges alone were not enough. Analysis by the Brennan Center for Justice showed that no company had fully kept its word, less than half provided the promised progress report, and those who did were poor on specifics. Public enforcement remained slow and regulation was increasingly absorbed into wider debates about simplification and competitiveness. At the same time, the scale of the technological shift continued to grow. Computing capacity has reportedly increased at extraordinary speed, and AI tools have become sufficiently prolific that they are now entering mainstream political and social communication rather than remaining confined to experimental or specialist use. From Slovakia to Pakistan, and from Argentina to India, documented cases have shown how AI can be used to distort campaigns, impersonate political actors, and manipulate the wider information environment.

Democratic procedures depend on more than formal voting rules. They require that citizens can participate in an information environment they reasonably trust, and be confident that fellow citizens are working from a recognisably shared body of knowledge. In electoral settings, this trust is especially fragile: the timeframe is short, the incentives to manipulate are high, and the cost of delay is often irrecoverable once voters have already formed impressions. AI, as currently deployed and insufficiently governed, is eroding these conditions at scale even as it promises lower costs, wider access to information, and more personalised participation. The International Panel on the Information Environment found that in 2024 alone, 80 per cent of countries holding national elections experienced documented AI-related incidents, and 69 per cent of those incidents were assessed as harmful. This is the baseline from which the next electoral cycle begins.

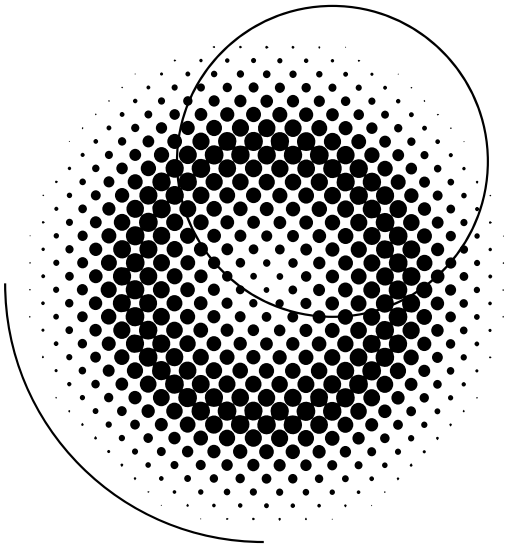
The distinctive challenge is that AI can produce inaccuracies and intentional falsehoods at the exact moment when democratic systems are least able to absorb them. Elections are unusually attractive targets because they compress attention,

reward speed, and create a high-value environment for strategic interference. Unlike other crises in digital democracy, electoral harms are often tightly timed, deliberately targeted, and capable of generating durable effects before correction can take place. A deceptive message, synthetic voice, or chatbot error may last only hours, but that may be long enough to shape a vote. The result is a distortion of democratic competition that can accumulate power in the hands of a narrow set of actors while weakening the public's ability to contest what it sees.

This report therefore proposes a citizen-centred framing of AI's impact on elections. It is designed to capture the full range of direct and indirect interactions that shape citizens' journey to decision-making — the full 'Road to the Ballot' — rather than focusing only on isolated forms of content manipulation. Production, interaction, communication, and campaigning form a cycle, and the report approaches them accordingly. The framework is global in scope, while remaining attentive to the different ways AI may affect electoral processes in different contexts.

The report also proposes standards that should apply at each phase of this cycle. The aim is not to regulate AI in the abstract, but to identify where electoral integrity is most exposed, where existing frameworks are incomplete, and where clear obligations are needed to preserve trust, fairness, and contestability. If that does not happen, the likely outcome is not only more disinformation, but a deeper and more durable concentration of informational power in the hands of the AI systems providers.

2026 also marked the death of Jürgen Habermas, a giant of democratic thought. One of the central lessons of his work is that democracy is a volatile and complex system which has historically managed to generate unprecedented social prosperity and individual dignity, but only under a rare balance among social actors. The political will required to sustain that balance has become increasingly difficult to maintain in the context of rapidly advancing technology. If left unchecked, AI revolution will bring about a distortion of democracy in which capital, data, and technological capability become increasingly concentrated in the hands of a narrow few, creating a deeper structural threat than the more familiar crisis of social media misinformation alone. This report seeks to provide a map for preserving the balance that has enabled previous generations of democratic societies to flourish.



Selected Incidents

More than four billion people participated in elections globally over the last four years. AI-enabled interference in electoral environments was documented in democracies at every stage of institutional development. The cases below illustrate that AI's impact on electoral integrity does not operate through a single mechanism. None of these can be adequately addressed by intervening at any other stage alone.

While politicians globally have primarily used artificial intelligence for campaign messaging, there are documented instances in which external actors have employed AI-generated content to manipulate public discourse by creating fraudulent voicemails or fabricated telephone recordings between candidates and prominent figures. In Nigeria, for example, a national news outlet disseminated a recording allegedly capturing a conversation between a presidential candidate and a religious leader; the candidate subsequently denounced it. Such instances are often more subtle and may receive less public scrutiny than the direct application of AI by official political campaigns. Consequently, initiatives by social media platforms and regulatory bodies to mitigate AI-generated disinformation must address the activities of both political campaigns and external entities that may more easily evade detection. In cases where media platforms themselves may be directly involved in generating and disseminating AI-generated or altered content without sufficient disclaimers, regulatory bodies must also establish robust oversight mechanisms to ensure that media consumers are notified of these cases, that platforms label AI-generated or altered content, and that victims of targeted disinformation campaigns receive sufficient redress.

Chinasa T. Okolo
AI Researcher, Strategist, and Policy Advisor. Founder and Scientific Director at Technēcultură.

Slovakia. Two days before the September 2023 parliamentary elections, a synthetic audio recording circulated online that appeared to feature a leading opposition figure discussing vote-buying. The timing exploited the campaign blackout period, while no electoral-specific safeguard had been put in place to prevent or quickly contain the harm.

Indonesia. During the February 2024 general election, a deepfake video of the deceased former president Suharto was circulated as campaign material and amplified across major platforms. The main failure lay in the speed and structure of platform amplification, which allowed it to reach large audiences before effective moderation could take effect.

Pakistan. In the 2024 election period, synthetic audio and video featuring ex-Prime Minister Imran Khan were circulated while he remained imprisoned and barred from campaigning. The material functioned both as impersonation and as a tool for mobilisation, in an environment where no binding framework existed to govern such use.

Romania. The first round of the 2024 presidential election was later annulled after coordinated manipulation of the information environment, including rapid amplification of content on TikTok. The central failure was one of disclosure: by the time the scale and nature of the distortion became clear, the result had already been produced.

Kenya. During the 2022 general election, AI-assisted political messaging was used across digital platforms in ways that showed both the opportunities and the risks of emerging campaign technologies. This illustrates the need for a coordinated response, combining state authority, international resources, and the ethical standards of non-governmental entities to safeguard against targeted polarisation during electoral processes.

India. The 2024 general election saw a surge in AI-based political content, including deepfakes and manipulated campaign material circulated at scale. The size of the electorate made the effect especially significant, as synthetic content could move rapidly across platforms before it was clearly challenged.

Argentina. Before the 2025 Argentine legislative elections, a fake AI video circulated claiming former president Macri had abandoned candidate Silvia Lospennato to back Javier Milei's spokesperson Manuel Adorni. Platforms removed the video and issued corrections, demonstrating escalation and notification duties to limit the impact of synthetic content in the final phase of an election.

Ecuador. During the 2025 presidential election, concerns were raised about the use of AI in the dissemination of disinformation and manipulated political content. This example shows that the problem is no longer confined to a single region or electoral model, but has become part of broader information-integrity risks.

Rwanda. During Rwanda's 2024 presidential election, researchers identified an AI-driven propaganda operation that used synthetic images and text to generate and spread messaging in support of President Kagame at scale. The campaign blended automated content creation with coordinated distribution across social media, mimicking grassroots support and inserting pro-government narratives into public discussion.

United States. In January 2024, a synthetic robocall impersonating President Biden sought to deter voters from participating in the New Hampshire primary. Independent analysis linked the voice clone to ElevenLabs technology, after which the company suspended the responsible account and later strengthened its safety and voice-protection measures, illustrating both the electoral risk of consumer voice cloning and the limits of reactive provider safeguards.

Canada. In 2025, election authorities and security officials warned that AI could amplify phishing, impersonation, and other attacks on democratic processes, including through deceptive communications directed at officials and voters. The case is important because it broadens the map from misinformation to the security of electoral institutions and personnel.

OpenAI / global. In its published threat reporting, OpenAI stated that threat actors had used its models for phishing, malware development, influence operations, and related operational support, indicating that AI is already being used to scale both cyber and information threats. Notably, this evidence derives from provider-reported misuse rather than from external observation alone.

Google / global. Google's published work on generative AI misuse and subsequent threat-intelligence reporting described the use of AI tools for impersonation, deceptive content, phishing, and malware support, illustrating the growing convergence of cyber risk and influence operations. As with the OpenAI case, the relevance of this example lies in the fact that the provider itself has documented the pattern of misuse.

Anthropic / global. Anthropic's safety reporting highlighted strategic deception and other harmful model behaviours under pressure, underscoring the need for stronger safeguards around frontier systems even where no single electoral incident is at issue. The case serves as a capability warning: systems deployed into democratic information environments may exhibit forms of manipulation or misalignment that existing governance arrangements are not yet equipped to address.

Grok / global. Reporting on X's chatbot documented the generation and spread of inaccurate election information, with election officials required to respond in real time to correct false claims. What makes this example significant is that the distortion occurred at the point of query, without the need for a viral deepfake or an immediately visible campaign artefact.

AI-mediated electoral queries / global. During the 2024 electoral super-cycle, LLMs embedded in high-traffic platforms, including Grok on X and Google's AI Overview, generated inaccurate responses to queries about candidates, polling dates, and results. CETaS research found that AI-generated content and influence operations could still "influence election discourse, amplify harmful narratives and entrench political polarisation," even where there was no conclusive evidence of a changed result.

The Road to the Ballot Framework

The framework is grounded in the observation that citizens do not encounter AI as a single, discrete object, but as a sequence of mediated experiences that shape how they perceive information, assess credibility, and act politically. From the perspective of user-centred design and anthropology alike, the meaning of a system is produced through its use. In electoral settings, this is particularly significant, as the same tool may be experienced differently when it generates content, answers a query, distributes information, or assists with a campaign. A framework built around the citizen therefore allows the analysis to proceed from lived exposure rather than solely from institutional compartmentalisation.

This approach is also intended to address a recurring limitation in existing governance models. Regulatory responses have often been organised by sector, by technology, or by actor, whereas the harms relevant to elections tend to cross those boundaries and accumulate across time. A citizen-centred framework makes it possible to identify where a given risk begins, where it is amplified, and where it becomes visible to the voter, while also preserving the distinctions necessary for effective policy design.

The Road to Ballot framework identifies four stages:

① **Production**

② **Interaction**

③ **Communication**

④ **Knowledgeable choice**

① Production

Production refers to what AI creates before citizens encounter it, including synthetic text, audio, images, and video, as well as personalised content generated for electoral purposes. At this stage, the central concern is that material may be fabricated or selectively generated at scale while appearing authentic, thereby weakening the ability of citizens to distinguish between verified content and synthetic output.

The dangers include:

① The creation of false or misleading political material.

② The mass production of content that is difficult to attribute, trace, or verify.

③ The absence of disclosure, watermarking, or provenance safeguards.

④ The use of AI to generate content designed to exploit electoral sensitivities or imitate identifiable political actors.

The appropriate policy response is to strengthen recognisability and traceability, including through provenance standards, disclosure obligations, and targeted restrictions on high-risk electoral use.

② Interaction

Interaction refers to the way AI responds to citizens' queries, including search tools, chatbots, recommendation systems, and other interfaces that mediate access to electoral information. The principal risk is not only that the answer may be inaccurate, but that it may be delivered in a form that appears authoritative. Sycophancy bias means that, in seeking to satisfy the person asking, these tools can weaken common factual reference points. Illusion of certainty refers to the fact that immediate answers appear neutral, increasing the practical effect of error.

The dangers include:

① Incorrect or misleading responses to questions about candidates, voting procedures, or results.

② The presentation of uncertain or disputed information with unwarranted confidence.

③ Bias embedded in ranking, source preference, or recommendation functions.

④ The absence of electoral-specific testing, auditing, or correction mechanisms.

The policy objective is reliability under electoral conditions, which requires heightened accuracy testing, transparent correction obligations, and constraints on electoral information functions where confidence thresholds cannot be met.

③ Communication

Communication refers to the way AI reshapes the exchange of information and opinion between citizens, parties, media actors, and institutions. The central issue is the transformation of the shared public sphere, including the amplification of synthetic content and the segmentation of audiences. These effects may be diffuse, but they can nonetheless alter the conditions under which citizens deliberate and form judgments.

The dangers include:

① Amplification of misleading or manipulated content.

② Fragmentation of the information environment into separate and unequal audiences.

③ Poor quality of automated moderation without human oversight.

④ The erosion of shared factual baselines across the electorate.

The policy objective is openness with accountability, which points to platform transparency, monitoring of algorithmic amplification, and timely public disclosure where information integrity is materially affected.

④ Knowledgeable choice

Knowledgeable choice refers to the use of AI to influence electoral participation directly, including voter persuasion, mobilisation, impersonation, and microtargeting. The concern is that AI may intensify existing asymmetries in campaign competition while enabling forms of influence that are difficult to detect, regulate, or contest, thereby affecting both the fairness of competition and the freedom of voter choice.

The dangers include:

① Deceptive or manipulative campaign content generated at scale.

② Advertising targeting that exploits social, ethnic, or demographic vulnerabilities.

③ The impersonation of candidates, parties, or public figures.

④ The use of AI to shape participation through channels that are not visible to opponents or regulators.

These risks may overlap with production, interaction, and communication, and in practice the same campaign activity may traverse all three before reaching the voter. The policy objective is proportionality and contestability, which requires disclosure, limits on deceptive practices, and rules that preserve the ability of opponents, regulators, and the public to scrutinise the use of AI in campaign communication.

Standard Proposals by Phase

Scope and principles

The Road to Ballot framework organises governance obligations around four stages of AI's intersection with democratic participation. This section sets out what those obligations should look like in practice. Each proposal is structured around a governance gap that existing instruments do not yet address, and a standard designed to close it. The proposals are intended to be precise enough to serve as reference standards for institutional adoption, while remaining technology-agnostic enough to remain valid as AI capabilities continue to evolve.

The proposals are not exhaustive. They identify the minimum set of standards at each stage without which the framework cannot function as a coherent whole. Additional standards, and the detailed technical specifications they require, are matters for the expert and institutional processes this report is intended to inform. The standards are framed as global reference points, adaptable to different legal systems and electoral contexts, but grounded in the common requirement that electoral governance preserve equal participation, public trust, and the integrity of the information environment.

Stage 1 — Production: What Synthetic Content is Created?

The production stage addresses AI systems at the point of content generation, before that content enters the information environment, reaches platforms, or is encountered by citizens. The governance gap at this stage is not the absence of any regulation, but the absence of requirements calibrated to the specific conditions of the electoral environment, where the timing, targeting, and format of synthetic content determine its harm potential as much as its content does.

Standard P-1: Differentiated mandatory labelling of AI-generated political content.

Existing labelling obligations do not differentiate whether content has political relevance or not. A deepfake video of a candidate requires a materially different disclosure (in terms of prominence, format, and placement) than an AI-assisted press release. P-1 proposes binding, content-type-specific labelling requirements for AI-generated or substantially AI-modified political content, applicable during defined electoral periods, designed for citizen legibility rather than technical compliance, and tested against the demonstrated ability of a non-specialist user to identify, understand, and act on the disclosure. Liability must be attached to both the person who removes or obfuscates the label and the provider of the AI system who did not mark the content in the first place.

A common international labelling taxonomy is necessary to prevent fragmentation of disclosure standards across jurisdictions, which currently allows content produced under weaker standards in one place to circulate without adequate disclosure elsewhere. Electoral authorities should be required to publish plain-language public guidance on what AI labels, updated before each electoral cycle, and

minimum visual prominence standards should be specified for video and audio content, where the current absence of such standards produces labelling that is technically present but practically invisible. Particular attention is required for AI-generated content in minority and regional languages, where labelling infrastructure is often weakest and where the populations most vulnerable to information environment manipulation are frequently concentrated.

Standard P-2: Watermarking and provenance standards.

The Coalition for Content Provenance and Authenticity framework offers a technically credible mechanism for content provenance tracking. It has not yet been adopted as a binding standard in electoral governance. P-2 proposes mandatory provenance metadata for AI-generated political content produced by regulated providers, preserved through the full distribution chain, with platform obligations to surface that provenance data in a citizen-accessible format at the point of encounter, not only in technical metadata readable by specialists and machines. Binding requirements on AI providers to maintain content provenance infrastructure should be treated as a condition of operating in electoral contexts, not as a voluntary commitment reviewable at the provider's discretion. Interoperability requirements between provenance systems are necessary to prevent platform-specific fragmentation from undermining the standard's coherence. Provenance infrastructure should also be privacy-preserving, using only the minimum metadata needed for provenance and identification of the deployer in case of incident occurrence, but not exposing user identity or sensitive personal data. It should also be adversarially robust, meaning resistant to removal or forgery under defined attack conditions. International alignment with existing provenance and standards processes is necessary so that standards developed in one jurisdiction can remain effective across borders.

Standard P-3: Platform detection and reporting obligations.

P-3 addresses platforms' obligation to detect AI-generated political content and AI-driven networks circulating on their systems during electoral periods. It proposes mandatory deployment of detection systems during defined electoral periods, with minimum accuracy thresholds set by the competent authority rather than self-determined by platforms, and periodic reporting to electoral and regulatory authorities on volumes, types, and reach of detected AI-generated political content, disaggregated by language and content type, not only in aggregate figures that obscure the uneven distribution of harms across linguistic communities.

Public transparency reports should be published on a defined schedule, with additional reporting obligations for coordinated inauthentic behaviour involving AI-generated content, including multi-agent amplification networks documented in the academic literature on electoral AI threats. During the final 72 hours before polling (the window most acutely vulnerable to fabricated content) real-time reporting obligations should apply to high-reach AI-generated content, with independent audit rights available to competent authorities over platform detection systems and whistleblower protections for platform employees who report failures in detection or reporting processes.

The Road to Ballot framework places emerging AI transparency and disclosure obligations within the context of the core institution of democracy: elections.

Having actively shaped implementation guidelines for Article 50 of the EU AI Act, I am encouraged to see transparency practices translated into a framework for an area of clear public-interest significance. Elections have perhaps been among the domains most harmed by coordinated misinformation and manipulation campaigns, and with Article 50 obligations becoming applicable soon, we can hope to see real improvements in our online information environment.

Turning to Production standards, P-1 recognises the importance of disclosure requirements for any AI-generated political content, in line with Article 50, which applies not only to professional actors but also to individuals generating AI content on matters of public interest, bringing bad actors into scope.

P-2 tackles the challenge of interoperability: viral political content migrates rapidly from VLOPs into encrypted environments like Telegram. Interoperable labelling and provenance measures are essential to enforce accountability across the ecosystem.

P-3 proposes currently absent monitoring and incident-reporting mechanisms capable of identifying unlabelled deepfakes and AI-generated political content at scale, providing authorities, platforms, and citizens with visibility into platform responses and bot-driven amplification during electoral periods.

Marta Bieńkiewicz
Policy and Partnerships Manager.
Cooperative AI Foundation

Stage 2 — Interaction: When Citizens Seek Information from AI

The interaction stage addresses AI systems at the point of direct citizen engagement, principally large language models and AI-assisted search and information systems through which voters increasingly seek electoral information. The governance gap here is substantial: no existing regulation imposes accuracy obligations on such systems specifically in the context of electoral queries, and no binding standard governs the training data practices that may shape the political dispositions those systems express.

Standard I-1: Accuracy obligations for LLMs in electoral contexts.

I-1 proposes a defined category of "electorally sensitive queries" covering candidates, parties, polling logistics, electoral results, and electoral processes to which heightened accuracy obligations apply during electoral periods. LLM providers operating systems that receive electorally sensitive queries should disclose known limitations at the point of interaction and provide a clear prompt directing users to authoritative electoral sources when the system cannot verify the answer with sufficient confidence. Providers should maintain and publish electoral-period error logs, available to competent authorities, and where an LLM has been found to have provided inaccurate electoral information at scale, affected users should be notified. Safe-harbour provisions should be available for LLMs that proactively redirect electoral queries to authoritative sources (electoral authority websites, official candidate registers) provided those redirections are consistently applied and themselves subject to accuracy verification. Particular attention is required for LLMs integrated directly into social media platforms

and search engines, where electoral query volumes are highest, user trust in the information environment is most easily exploited, and accuracy failures are most consequential at scale.

Standard I-2: Training data integrity and bias disclosure.

The political dispositions expressed by LLMs in response to electoral queries are not solely a function of real-time moderation. They reflect the composition of training data, an upstream governance issue that current frameworks almost entirely ignore. I-2 proposes mandatory documentation of training data sources for general-purpose AI models deployed in electoral contexts, available to competent authorities, with binding obligations to identify and remediate training data that has been subject to coordinated manipulation designed to skew electoral outputs. Deliberate seeding of web-scale training corpora with disinformation represents a structural vulnerability in LLM integrity that labelling and accuracy requirements at the output stage cannot address alone. Periodic bias audits of LLM outputs on politically sensitive topics, conducted by independent auditors with access to model internals, should be a condition of deployment in electoral contexts, with public disclosure of known political biases identified in model outputs updated at minimum annually. A research mandate for investment in detection methodologies for training data poisoning (currently a significant technical gap) should accompany the disclosure requirements.

Stage 3 — Communication: How AI Shapes What Citizens See

The communication stage addresses AI systems that mediate the distribution of political information to citizens, principally recommendation and content moderation systems operated by large platforms. The governance gap at this stage is only partially addressed by existing legal instruments, but these instruments were not designed as electoral standards, and their enforcement timelines, audit mechanisms, and transparency requirements are often calibrated to the pace of ordinary regulatory procedure rather than to the pace of the electoral calendar.

Standard C-1: Transparency and auditability of AI recommendation systems.

C-1 proposes mandatory disclosure to competent authorities of the parameters governing AI recommendation systems in contexts with electoral relevance, including the weighting given to political content, the criteria for content promotion or demotion, and any modifications to those parameters during the electoral period. Such disclosure should be made to electoral authorities on a confidential basis, with appropriate safeguards for trade secrets, rather than published publicly.Independent audit rights over recommendation systems during defined electoral periods, exercisable by national electoral authorities and coordinated internationally, should be available as a matter of legal obligation, not platform discretion. Citizens should have access to a clear, plain-language explanation of the primary factors determining the political content they are shown, not the full technical specification, but a genuine account of the system's logic, updated in real time or at minimum per electoral cycle. Critically, recommendation system modifications during the final 30 days

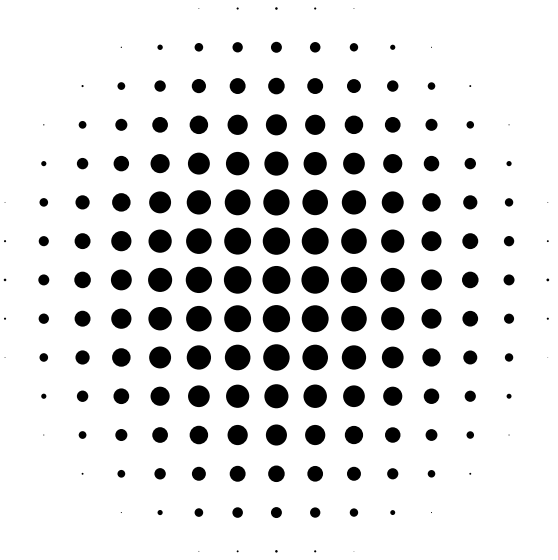
before polling should be prohibited unless disclosed in advance to competent authorities, a provision that directly addresses the capacity for platforms to alter the information environment in electoral-period-specific ways without public accountability. Post-election audit access to logs of content promotion and demotion decisions made during the electoral period should be a standard obligation, alongside mandated data access for accredited academic researchers studying recommendation system effects on electoral perception.

Standard C-2: Language-disaggregated moderation quality requirements.

C-2 is the equity standard of the framework. Aggregate platform-wide moderation accuracy figures conceal an unequal distribution of protection, systematically skewed against smaller language communities, precisely the communities most vulnerable to information environment manipulation and least resourced to identify and contest moderation failures. C-2 proposes binding minimum accuracy requirements for AI-supported content moderation of political content, disaggregated by language, with mandatory human review escalation pathways for political content moderation decisions in languages where automated accuracy falls below defined thresholds. Transparency reporting on moderation accuracy by language should be published at minimum annually and during each electoral period. Every moderation decision affecting political content during an electoral period should be subject to a right of appeal to human review, regardless of the language in which the content appears, with a defined maximum response time calibrated to the electoral calendar. Specific provisions are required for languages spoken across multiple jurisdictions, where cross-border moderation inconsistency is documented and where the same content may receive different treatment depending on the national context in which it is encountered.

Standard C-3: Advertising with the use of AI.

The emergence of advertising inside LLM responses, as well as increasingly automated targeting systems on major platforms, creates a new governance issue at the intersection of information and promotion. Where ads are inserted into conversational interfaces, they should be clearly separated from organic responses and explicitly labelled as sponsored content, with safeguards to prevent interference with the quality or neutrality of the answer. Where platforms rely on automated targeting, including systems that infer audience characteristics or match ads to user behaviour, the criteria for delivery should be subject to transparency and review, especially in electoral contexts. The standard should not prohibit advertising, but it should ensure that users can distinguish between information, recommendation, and promotion, and that political ad delivery remains subject to meaningful oversight rather than opaque automation.



Stage 4 — Knowledgeable Choice: Protections at the Moment of Electoral Participation

The campaign stage addresses the point at which political actors make decisions about how to deploy AI tools in their electoral communications, the governance gap closest to the ballot itself. The Slovak, Indonesian, Romanian, and Kenyan cases all involved, at some point, a decision by a political actor or their proxies to produce or distribute AI-generated or AI-amplified content in an electoral context. Standards at the production, interaction, and communication stages address the infrastructure through which that content operates; campaign-stage standards address the actors who deploy it.

Standard K-1: Mandatory disclosure of AI use in political communications.

K-1 proposes binding disclosure requirements for the use of AI in the production or targeting of political communications, applicable to all actors subject to electoral law: parties, candidates, campaigns, and third-party political advertisers. Disclosure obligations should extend to AI-assisted microtargeting and audience segmentation, not only to AI-generated content: a political communication produced by a human but targeted through AI-driven behavioural profiling is as much a product of AI deployment as a fully synthetic video. Disclosure requirements should be integrated into existing political advertising transparency registers, where such registers exist, and otherwise embedded in national electoral disclosure rules. Enforcement authority should vest in electoral authorities, with sanctions proportionate to reach and electoral impact, and disclosure information should be archived post-election and made available for research and audit purposes. Particular attention is required for AI-generated social media content produced by campaign staff using consumer AI tools, a category that is currently outside most disclosure frameworks and that represents a rapidly expanding vector of AI use in campaign operations.

Standard K-2: Prohibition on synthetic impersonation of candidates.

K-2 proposes a binding prohibition on the creation or distribution of AI-generated content that synthetically reproduces the voice, likeness, or stated positions of a candidate or electoral official without their explicit, documented consent. Non-consensual synthetic impersonation of candidates during electoral periods should be classified as a category of electoral fraud, subject to electoral law rather than only content moderation policy, a classification that attaches both greater legal consequence and clearer enforcement authority than treatment as a platform terms-of-service violation. Platform obligations to remove prohibited impersonation content within defined timeframes during electoral periods, with liability for non-compliance, should accompany the prohibition. Industry standards analogous to voice protection registries (mechanisms allowing individuals to register their voice against unauthorised synthetic reproduction) should be mandated across major AI audio and video generation providers rather than left as voluntary and unevenly adopted. Candidates who have been subject to synthetic impersonation should have access to rapid injunctive relief, not only post-election redress: the temporal dynamics of electoral harm require that remedies be available on timescales compatible with the electoral calendar.

Standard K-3: Election incident protocols.

K-3 addresses the most significant structural gap in current electoral governance: the absence of a binding framework for the notification of citizens,

candidates, and electoral authorities when AI-driven threats are assessed as sufficiently severe to undermine electoral integrity. The question of who is responsible for issuing such notification, on what threshold, in what format, and within what timeframe is currently unanswered in most major governance instruments. K-3 proposes a formalised election incident protocol establishing the threshold criteria that trigger public notification, the institutional actors with authority to issue it, the speed and format of communication, and the obligations of AI providers to support timely disclosure. The protocol is modelled in part on existing public notification frameworks for election threats, adapted for AI-specific threat categories and diverse institutional contexts. Each state should designate, in advance of each electoral period, a competent authority with responsibility for assessing and triggering incident protocols, coordinated internationally through relevant electoral cooperation mechanisms. Binding obligations on AI providers to notify competent authorities of detected threats meeting defined severity thresholds, within defined timeframes, should accompany the institutional architecture. A graduated notification structure, distinguishing between authority notification, platform action, and public communication, is preferable to a binary trigger, both to prevent under-reaction to serious threats and to avoid alarmism where a lower-threshold public notification requirement could be counterproductive. Each triggered protocol should be followed by a published post-incident review assessing its adequacy and informing future threshold calibration.

As accessible AI tools expand the threat surface against election integrity, many countries are struggling to understand when or how to relay information to voters on how to protect themselves.

Relying on ad hoc decision-making during highly constrained election periods is unlikely to strike the right balance. When authorities fail to raise awareness quickly, public confidence in the robustness of democracy itself is undermined. On the other hand, overreporting on non-critical cases creates its own challenges, inducing "warning fatigue" and risks fuelling counterproductive alarmism over the threat from AI.

To break this gridlock, states must replace reactive improvisation with a formalized election incident protocol. By implementing structured procedures, authorities can accurately establish clear thresholds for public communication, systematically coordinate with AI companies, and maintain public trust at a time when state capacity is heavily strained.

Sam Stockwell
Senior Research Associate at the Centre for Emerging Technology and Security (CETaS) within the Alan Turing Institute.

Towards a Responsible Future

The democratic response to the changes brought about by the emergence of AI will fall short if it is reduced to the management of individual incidents, however serious those incidents may be. This report aims to demonstrate that electoral harm arises not only when a deepfake goes viral, a platform fails to moderate content, or a campaign deploys a deceptive tool. Where institutions respond in a fragmented manner, addressing one point of failure while leaving broader conditions of trust unresolved, public confidence in democracy is weakened. A governance approach that remains purely reactive will therefore struggle to preserve electoral integrity, because by the time the most visible harms appear, the conditions that enabled them may already have affected other parts of the system.

Realistically, not every instance of AI-related distortion can be prevented in advance. Such a standard would be unrealistic and, if pursued without restraint, could itself justify forms of intervention inconsistent with democratic freedoms. The relevant benchmark of success is whether the information environment remains sufficiently intelligible, contestable, and trustworthy for citizens to make political judgements with reasonable confidence, and whether institutions can respond in a timely, proportionate, and credible manner when those conditions are placed under strain.

Incident response remains necessary, but it cannot serve as the sole organising principle of electoral governance in the age of AI. As electoral harm is often cumulative, it crosses from content production to dissemination, interaction, and campaign deployment before it becomes visible in a form institutions can address. By the time a harmful item is identified and removed, its effects may already have been amplified, internalised, or redirected through other channels.

This is particularly important in electoral settings, where timing is itself part of the harm. A correction issued after voting has begun, or after a false claim has shaped public perception, does not restore the original conditions of fairness. For that reason, governance should be designed not only to react to incidents, but also to reduce the likelihood that predictable failures will cascade across the wider information environment. A resilient electoral information environment requires the preservation of conditions under which citizens can still recognise political communication, assess competing claims, and form judgments without being systematically disadvantaged by opacity, deception, or unequal access to reliable information.

Several features support that resilience: clarity, the ability to challenge decisions and information, institutional credibility, and a shared basis for public reasoning. Citizens should be able to understand when AI is involved, to challenge decisions and content that affect them, and to rely on the existence of institutions capable of responding in ways that are visible and credible. This has important implications for governance design. The relevant question is whether technical safeguards are legible to citizens and capable of sustaining public confidence across the electoral cycle. Labelling, provenance systems, moderation tools, audit mechanisms, and notification procedures are necessary parts of this architecture, but they work only when citizens can understand what they do, when they apply, and how to contest their failure.

At the same time, standards should remain proportionate. Measures adopted in the name of electoral integrity should not become so broad,

intrusive, or discretionary that they undermine political freedom, pluralism, or legitimate expression. Effective governance therefore depends on a balance: standards should be clear enough to constrain harmful conduct, but limited enough to remain compatible with democratic rights and the ordinary uncertainties of political life.

Implementation of these norms requires electoral authorities, regulators, courts, and other competent bodies to establish cooperation that goes beyond the formal allocation of responsibility. Separate bodies will fail collectively if they cannot provide timely guidance, intelligible public communication, technical access, and remedies calibrated to the pace of the electoral calendar. The ability to make governance visible, comprehensible, and credible from the perspective of the citizen is a key component of the institutional capacities to be built by public administration.

The same logic applies to platforms, AI providers, observers, and researchers. Platforms and providers should be able to act at the speed required by electoral harms, while independent observers need sufficient access to assess whether the safeguards in place are functioning in practice. International co-operation is equally important, since AI systems, platform infrastructures, and influence techniques do not remain confined within national borders, whereas many electoral safeguards still do.

The final test of any framework is what citizens actually see on their screens and how they respond to it. For that reason, implementation should include mechanisms for testing outcomes at the level of public experience: whether disclosures are noticed and understood, whether official corrections reach the audiences exposed to falsehoods, whether citizens can distinguish manipulated from authentic content with reasonable confidence, and whether institutional responses are regarded as timely and credible. These questions should be assessed through post-election reviews, independent audits, public survey research, and structured observation by electoral monitors, so that future standards can be evaluated by their practical effect on democratic choice.

Conclusions: Who Acts and How?

The principles outlined above will make a difference only once they are translated into concrete responsibilities for different actors. Protecting electoral integrity therefore requires a distribution of responsibility across institutions, companies, political actors, and oversight mechanisms, with each acting within a defined mandate and on an appropriate timescale. The following bodies are therefore called upon to act accordingly and to assume their respective share of responsibility:

States and legislatures bear primary responsibility for establishing the legal basis on which the framework set out in this report can operate. They should define the minimum rules applicable to AI-generated political content, electorally sensitive AI systems, platform transparency, and campaign conduct, and ensure that those rules are enforceable through electoral law, administrative law, or other appropriate legal instruments. Where the law remains silent or outdated, the practical effect is to leave key conditions of democratic participation to private incentives and uneven cross-border pressures.

Promising example: EU AI Act and Digital Services Act
The European Union has adopted the AI Act and the Digital Services Act (DSA), which together impose transparency, risk management, and accountability obligations on AI systems and online platforms, including during elections. These laws require labeling of certain AI-generated content and mandate platform transparency and risk mitigation, creating an enforceable legal baseline across member states.

Electoral authorities should serve as the operational focal point during electoral periods. Their responsibilities include issuing public guidance, interpreting campaign rules as they apply to AI use, co-ordinating incident response, and maintaining channels through which citizens can receive timely and authoritative information. Where they lack the mandate, expertise, or technical access required to perform these functions, states should close those gaps before the next electoral cycle begins.

Promising example: India Election Commission AI advisories
Ahead of the 2024 general election, the Election Commission of India issued formal guidance to political parties on the use of AI-generated content, including deepfakes. It clarified that existing campaign and misinformation rules apply to AI use and required pre-certification for certain political materials, demonstrating an operational role during the electoral period.

Platforms and AI providers are responsible for the systems through which political content is generated, ranked, distributed, recommended, and presented to users. In electoral contexts, this responsibility should take the form of enforceable public-interest duties, including labelling, provenance, detection, transparency, escalation, and notification mechanisms. Internal policies and voluntary commitments may support this work, but they cannot replace obligations that are externally defined, independently reviewable, and capable of enforcement.

Promising example: YouTube's 2026 AI labeling and automatic detection
In May 2026, YouTube introduced a unified, more visible label for photorealistic AI-generated or altered content, showing it directly below long-form playersand as an overlay on Shorts. The platform also began automatically applying these labels when its detection systems identify significant photorealistic AI, even if creators do not disclose, creating an enforceable mechanism for transparency and provenance in political and electoral contexts.

Political parties, candidates, and campaign organisations remain responsible for how AI is used in electoral competition. The use of synthetic media, behavioural targeting, automated persuasion, or other AI-assisted techniques does not place conduct outside electoral law or electoral ethics. On the contrary, the capacity of these tools to intensify asymmetries, obscure authorship, and manipulate voter perception makes transparent and lawful conduct more important, not less.

Promising example: South Korea deepfake election law
South Korea amended its Public Official Election Act to ban the use of deepfake videos in political campaigns within 90 days of an election. This directly places responsibility on candidates and campaign organizations, reinforcing that AI-assisted tactics remain subject to electoral law and ethical standards.

Regulators, courts, and oversight bodies are responsible for supervision and redress. Their role is to interpret standards, review compliance, adjudicate disputes, and ensure that enforcement remains consistent with due process and democratic freedoms. The credibility of the framework depends in part on their ability to act with consistency, legal clarity, and restraint.

Promising example: France's ARCOM supervision of online platforms
France's media regulator ARCOM oversees platform compliance with obligations related to misinformation and electoral integrity, including during election periods. It can issue formal notices and sanctions, illustrating how a regulatory body enforces standards and ensures proportional, legally grounded responses.

Civil society, journalists, and researchers provide scrutiny that neither governments nor companies can perform alone. They document harms, test systems, identify emerging practices, and explain complex developments to the public. A credible governance framework should therefore protect independent research access, support timely data availability, and avoid obstructing public-interest scrutiny of electoral AI systems.

Promising example: EU DisinfoLab investigations into influence operations
EU DisinfoLab conducts independent investigations into coordinated disinformation campaigns, including those involving synthetic or manipulated content. Its findings expose cross-platform influence networks and provide publicly accessible evidence that supports accountability and informed public debate.

International organisations, standards bodies, and intergovernmental forums have a co-ordinating role. AI systems, platform infrastructures, and influence operations routinely cross borders, while legal and electoral safeguards often remain national. Greater alignment in terminology, disclosure expectations, audit practices, and incident protocols would reduce avoidable gaps and make minimum protections more durable across jurisdictions.

Promising example: OECD AI principles and G7 Hiroshima AI Process
The OECD AI Principles and the G7 Hiroshima AI Process establish shared frameworks for trustworthy AI, including commitments relevant to information integrity and democratic processes. These initiatives promote alignment on transparency, accountability, and risk mitigation across countries, helping address cross-border challenges in electoral AI use.

Most of the above measures can be adopted in the near term, including disclosure duties, impersonation prohibitions, public guidance, and incident protocols. Others, such as interoperable provenance infrastructure, bias auditing at scale, and deeper transparency for recommendation systems, require longer-term investment and institutional capacity. The immediate task is therefore to assign responsibility clearly, prioritise measures that can already be implemented, and begin building the capacities on which the rest of the framework depends. The integrity of future elections will depend on whether this is done early enough, consistently enough, and with sufficient clarity to preserve informed and equal democratic choice.

Bibliography

Alan Turing Institute, Centre for Emerging Technology and Security. (2024). *AI-enabled influence operations: The threat to the UK general election.*

Alan Turing Institute, Centre for Emerging Technology and Security. (2024). *AI-enabled influence operations: Safeguarding future elections.*

Anthropic. (2026). *Responsible scaling policy version 3.0.*

Anthropic. (2026). *Risk report: February 2026.*

Brennan Center for Justice. (2025). *Tech Companies Pledged to Protect Elections from AI — Here's How They Did.*

CNN. (2024, 6 December). *Romania's top court annuls presidential election result.*

Demos. (2026, 19 May). *Safeguarding UK elections in the world of LLMs and AI chatbots.*

DW. (2024, 4 December). *Did TikTok influence Romania's presidential election?*

ElevenLabs. (2024). *No-go voices.*

Google AdSense Help. (n.d.). *How ads are targeted to your site.*

Google DeepMind, Jigsaw, & Google.org. (2024). *Generative AI misuse: A taxonomy of tactics and insights from real-world data.*

Google Threat Intelligence Group. (2025). *Advances in threat actor usage of AI tools.*

International Panel on the Information Environment. (2025). *Global trend of widespread role of GenAI in national elections in 2024.*

Meta. (n.d.). *Advantage+ shopping campaigns / automatically targeted ads.*

MCDT Cambridge. (2026). *Data Not Found: Social Media Data Transparency for Information Integrity.*

Munich Security Conference. (2025, February 15). *AI elections accord.*

OpenAI. (2025). *Disrupting malicious uses of AI.*

OpenAI. (2025). *Testing ads in ChatGPT.*

The Guardian. (2024). *X's AI chatbot spread voter misinformation — and election officials fought back.*

The Verge. (2024). *Two Texas companies were behind the AI Joe Biden robocalls.*

Vox. (2024, 2 January). *2024 is the biggest global election year in history.*

Wired. (2024). *Mystery company linked to Biden robocall identified by New Hampshire Attorney General.*