WITNESS Public Comment on the Role of Generative AI in Elections

May 7, 2026

AI-generated content is increasingly shaping electoral environments not primarily through isolated instances of highly deceptive deepfakes, but through broader, systemic effects on information ecosystems, public trust, and political discourse. Drawing on over three years of leading the Deepfake Rapid Response Force (DRRF) at WITNESS—an organization dedicated to protecting reality in the age of AI—we have directly supported responses to deceptive AI across numerous electoral contexts, including in Ghana, India, Georgia, Bolivia, Mexico, the Philippines, Indonesia, and the United States. This cross-regional experience provides a unique, comparative perspective on how AI-generated content is being deployed in elections globally, revealing both recurring patterns and context-specific dynamics.

First, AI significantly lowers the cost and barrier to producing persuasive, multimodal political content at scale. Across contexts, this has enabled the rapid production and large-scale dissemination of emotionally resonant videos, synthetic audio, images, and text. These are often distributed across platforms to amplify narratives, harass candidates, or suppress participation. Women politicians are disproportionately targeted – for example, analysis by the American Sunlight Project found that one in six U.S. congresswomen had been targeted with non-consensual intimate imagery (NCII) deepfakes. Importantly, the most impactful content is often not highly sophisticated, but strategically effective: leveraging fear, identity, or political division rather than relying solely on convincing deception.

Second, AI-generated content contributes to a broader erosion of trust in the information environment. The proliferation of synthetic media makes it increasingly difficult for audiences to distinguish between authentic and manipulated content, while also enabling "plausible deniability," where genuine content is dismissed as AI-generated. DRRF findings indicate that approximately one-third of cases submitted for analysis involve authentic media falsely labeled as AI-generated (by political actors or confused audiences). For example, during Ghana's elections, a Vice-Presidential candidate claimed that an authentic recording of his campaign speech was doctored. Similarly, a viral video of Benjamin Netanyahu in a coffee shop was widely dismissed online as a deepfake despite lacking evidence of manipulation and despite available indicators confirming its authenticity. This dynamic undermines not only trust in individual pieces of content but also the broader ability of citizens, journalists, and institutions to rely on audiovisual evidence in democratic processes.

Third, AI enhances the scalability and subtlety of influence operations, including those aligned with geopolitical narratives. The Hungarian elections provide a clear illustration. The Russian invasion of Ukraine was a central campaign theme, and AI-generated content was used to reinforce related fears. One widely circulated video showed a young girl mourning her father, allegedly sent to fight in Ukraine. The video was posted by the official Fidesz party account with the caption: "This video may be an AI video, but war is truly horrific!", implying that opposition

leadership would lead Hungary into war. Investigations by Lakmusz identified additional similar videos posted on the Facebook page "Nem a mi háborúnk" ("Not our war"), depicting fathers and husbands being sent to die on the frontlines. These examples demonstrate how AI-generated content can shape public sentiment without relying on fully deceptive deepfakes, instead embedding itself within broader political messaging and emotionally charged narratives.

More broadly in Hungary, AI-generated content was widely used to amplify narratives and simulate authenticity. The TikTok account "BrüsszelÜzem" (Hungarian for "Brussels Operation") posted news-style videos featuring AI-generated anchors and experts commenting on opposition figure Péter Magyar, often paired with authentic footage taken out of context. Other content included simulated voter interviews and AI-generated "vox populi" videos portraying everyday citizens expressing political views. Additional videos depicted families mourning Hungarian soldiers or imagined future scenarios, such as projections of Hungary in 2050, suggesting that voting for the opposition would lead to national decline. AI-generated content also included endorsements or commentary from historical figures, celebrities, cartoon characters, and even anthropomorphized figures (e.g., animated vegetables or a monkey performing ASMR while delivering political messages). These formats shape perceptions even when audiences recognize them as synthetic.

At the same time, the role of online "political influencers" is evolving in ways that further amplify these dynamics. Increasingly, electoral discourse is shaped not only by individual pieces of content but by networks of content creators who produce high volumes of politically aligned material. Recent reporting has documented the emergence of AI-generated influencer personas, such as pro-Trump "MAGA" influencers who blend political messaging with lifestyle and identity-driven content to build large, engaged audiences. These accounts often leverage highly stylized and gendered portrayals (such as idealized or sexualized female personas) alongside relatability and ideological alignment to establish trust and influence, even when their synthetic nature is disclosed or suspected.

Crucially, these influencers do not operate in isolation. Clusters of accounts, sometimes coordinated, sometimes loosely aligned through shared incentives, can function as influence ecosystems, reinforcing narratives through repetition, cross-posting, and algorithmic amplification. This creates a persistent and distributed form of narrative shaping, where political ideas are normalized and embedded into everyday content streams. In electoral contexts, such dynamics can manufacture a sense of consensus, distort perceptions of public opinion, and influence voter attitudes without relying on any single piece of clearly deceptive or policy-violating content.

The Hungarian case also highlights the "blended" nature of the current information environment. Highly realistic synthetic media coexist with satire, parody, and low-quality "AI slop." For example, the channel "PitiPeti kalandjai" portrayed Magyar in exaggerated caricature form, while the TikTok page "BiteGeg" described its content as "satirical parody and creative, humorous sketches," yet also posted videos reinforcing political narratives about forced conscription in Ukraine and alleged opposition policies. This mixture of authentic, synthetic, satirical, and misleading content blurs the line between fact and fiction, increasing cognitive load

for audiences and making credibility assessments more difficult. Importantly, evidence suggests that labeling content as AI-generated does not eliminate its impact. Even when disclosures are present, as in the Fidesz example, content can continue to shape perceptions and reinforce narratives. This underscores the limitations of transparency measures when deployed in isolation, particularly in high-volume, fast-moving information environments.

Beyond Hungary, patterns observed across other electoral contexts point to additional risks. Audio deepfakes, particularly in the form of alleged "leaked" phone conversations, are an emerging concern. In 2025, DRRF analyzed suspected fake phone calls in relation to elections in Bolivia and Iraq. These cases illustrate known challenges of AI audio detection–factors, such as low quality and high compression of audio files, background noise, cross-talking, or lack of representative datasets, including diverse local languages and accents, significantly limit the effectiveness of AI detection tools. These technical shortcomings, combined with the general inability of the human ear to spot them, especially audio deepfakes that present a concern in the electoral setting, make it easy for political actors to discredit authentic audio recordings as deepfakes, contributing to the broader ecosystem of doubt.

Taken together, these dynamics have concrete implications for democratic processes. AI-generated content can suppress participation by increasing confusion and distrust, distorting perceptions of public opinion through manufactured consensus, and creating chilling effects for candidates, particularly women, who face disproportionate harassment. More broadly, it undermines the role of audiovisual evidence in journalism, accountability, and electoral dispute resolution.

For platforms, these developments pose significant governance challenges. Existing approaches, particularly those focused on labeling or detecting individual pieces of content, are insufficient to address the systemic and networked nature of AI-driven influence, as well as the real-world conditions under which such content is produced and shared (including low quality, compression, limited context, and multimodal complexity), as demonstrated in WITNESS' TRIED benchmark. Much of the most impactful content is not technically clearly marked as deceptive or violative, but operates through scale, repetition, and narrative reinforcement. Addressing these risks must require broader adoption of provenance solutions (such as C2PA), consistent and meaningful labeling practices, and greater attention to coordinated networks, cross-platform dynamics, and the known limitations of current detection tools, particularly in under-resourced languages and contexts.

Overall, the role of AI in elections is best understood not as a problem of isolated deceptive artifacts, but as a structural transformation of the information ecosystem. AI-generated content amplifies polarization, enables scalable and often opaque influence operations (including potential foreign interference), facilitates harassment and targeted attacks, and erodes the shared baseline of trust that democratic processes depend on. The Hungarian elections are one example within a broader global trend, where the cumulative effect of AI-generated and AI-amplified content reshapes how political narratives are produced, distributed, and believed.