CEE Digital Democracy Watch

Overseeing the New Digital Era in Election Integrity

Warsaw, 28.05.2026

Public Comment — Case 2026-034-FB-UA
AI Video Faking UK Politician's Immigration Views

Submitted by CEE Digital Democracy Watch

This comment is submitted in response to Case 2026-034-FB-UA, in which the Oversight Board examines an apparently AI-generated video impersonating a UK Labour Party councillor from Scotland and misrepresenting her views on immigration.

CEE Digital Democracy Watch has recently commented on AI-generated political content in the Hungarian elections and published analysis on AI-driven disinformation and synthetic harassment in Europe. Most risks identified there are directly relevant to this case. The case sits within a wider pattern of AI-generated divisive content, including overseas "content farms" and commercial networks monetising anti-migrant deepfakes aimed at Western audiences.

1. Non-labelled AI misrepresentation and divisive messaging

The core harm is the non-labelled use of AI to put inflammatory words into the mouth of a real politician in a racially charged context. The councillor publicly rejected racist disinformation about asylum seekers and supported housing them in her area; the AI video reverses this by fabricating a quote that trivialises rape and portrays her as endorsing a grotesque "refugees are welcome even if they rape our women" stance. This is not ordinary satire or criticism but a synthetic misrepresentation of a politician's beliefs on a highly sensitive issue.

Under the EU AI Act, such material squarely fits the notion of a "deepfake": AI-generated or manipulated content that appreciably resembles real persons and that a reasonable viewer could mistake for authentic. Article 50 AI Act requires that deepfakes be clearly labelled as artificially generated or manipulated, and that generative outputs be

technically marked as synthetic. Even though the UK is outside the EU, these rules express a broader democratic norm: citizens must be able to distinguish genuine political speech from AI-generated impersonations, especially on contentious issues like immigration and violence against women.

Leaving such content unlabeled contributes to epistemic manipulation: the distortion of the informational conditions under which people form beliefs about political actors and policies. Empirical work shows that vivid political deepfakes can influence perceptions even when audiences know deepfakes exist; later corrections often cannot fully undo the initial impact. Here, the synthetic quote is crafted to trigger anger, fear and moral outrage around rape and asylum, heating divisive emotions and skewing public opinion about the councillor and refugee housing.

Meta argues that the content did not require an AI label because the councillor was not on the ballot, the post was "satirical," engagement was low and there was no "crisis." This treats risk as a function of metrics and narrow temporal windows, rather than context, vulnerability and the nature of the misrepresentation. The duty to reduce deception about core political positions cannot be limited to viral content or formal election periods.

2. AI-enabled harassment of politicians and vulnerable groups

Meta's decision relies heavily on the councillor being an adult public figure "not protected from unwanted manipulated imagery" under its Bullying and Harassment policy. In an era of AI-enabled harassment, such a rigid public-figure distinction is no longer adequate. Local politicians, especially women and those speaking on contested issues such as migration or gender-based violence, face coordinated campaigns that can endanger their safety and drive them out of public life.

CEE Digital Democracy Watch's report *Non-Consensual Sexualising Deepfakes – Threats and Recommendations for Legal and Societal Action* shows that the overwhelming majority of deepfake videos online are sexualising in nature and that almost all known cases target women, including public figures. These attacks aim to silence or discredit women and cause severe psychological, social and professional harm. While the present case is not sexualised, it shares key features: the use of synthetic media to distort a woman politician's identity, exploit gendered fears ("rape our women") and expose her to heightened risk in an already hostile environment.

From a human rights perspective, public-figure status should not strip individuals of robust protection against AI-generated harassment and impersonation. Platforms should recognise that certain public figures, including women, racialised minorities and those defending migrants or other marginalised groups, are particularly vulnerable to AI-enabled attacks and warrant a higher standard of diligence. This includes a right to timely human review when they are impersonated with AI and a presumption in favour of protective action where safety concerns are reasonably raised.

In this case, two user reports and two appeals did not trigger human review, and the content was never fact-checked or escalated via Trusted Partners. Such gaps mean that AI-generated defamation of at-risk politicians can remain online by default, even when it clearly misrepresents their views on matters of intense public controversy.

3. Globalised AI content farms and systemic incentives

The UK case also reflects systemic incentives that encourage AI-generated divisive content. BBC and other investigations have documented overseas content farms in Vietnam and elsewhere producing AI-generated deepfake videos of UK politicians, which Meta removed only after media contact, while near-identical pages reappeared shortly thereafter. Parallel work has tracked large networks of anti-immigration AI content aimed at UK audiences back to operators in Sri Lanka and other countries, running dozens of pages that publish synthetic protest scenes, dystopian imagery and fabricated quotes to generate engagement and ad revenue.

One investigation found that a Sri Lankan influencer built a network of over 100 Facebook pages focused on UK anti-migrant narratives, using generative AI to produce content that attracted millions of views and substantial income; tutorials linked to this network actively encourage students to use AI to exploit UK immigration as a “strong trigger” for engagement. Similar dynamics have been observed in Central and Eastern Europe, where Russian-linked disinformation networks and domestic actors deploy AI-generated content and inauthentic accounts to inflame fears about war and refugees and undermine trust in institutions.

If AI-generated impersonation of politicians is tolerated without labelling or robust moderation, it signals to commercial operators and hostile states that such tactics remain a low-risk, high-reward instrument for shaping public debate, whether in the UK, Hungary or beyond.

Recommendations

In light of the above, CEE Digital Democracy Watch proposes that the Oversight Board consider the following recommendations to Meta:

1. **Address monetisation incentives for deceptive AI political content.** Review recommendation and monetisation systems to ensure that pages and accounts repeatedly using AI-generated misrepresentation and hate-baiting are excluded from monetisation and do not receive algorithmic amplification. Meta should cooperate with regulators and independent researchers to analyse the financial incentives driving such networks, including overseas content farms profiting from UK-focused anti-migrant deepfakes.
2. **Create a high-risk category for AI impersonations of political actors on divisive issues.** Explicitly classify realistic AI-generated or manipulated depictions of politicians and public figures as a high-risk category when they misrepresent views on sensitive topics such as immigration, race or gender-based violence. Such content should receive clear, prominent AI-generated labels at first exposure and robust technical marking (e.g. watermarking, metadata), reflecting the transparency and anti-manipulation principles of the EU AI Act and similar standards.
3. **Guarantee priority human review for AI-generated attacks on at-risk public figures.** Introduce a dedicated escalation pathway for reports involving AI-generated impersonation or harassment of politicians and activists working on controversial issues, including immigration and violence against women. Reports in this category should be reviewed by trained human moderators within defined timeframes, with a minority languages focus and a meaningful appeal process, rather than relying solely on automated triage or Trusted Partner flags.
4. **Strengthen protection for vulnerable groups, especially women, against AI-enabled harassment.** Recognise women politicians, journalists and activists as a group particularly exposed to AI-enabled abuse, given that over 90% of deepfake content is sexualising and overwhelmingly targets women. Commit in policy and enforcement to heightened scrutiny and rapid intervention when AI-generated content weaponises gendered tropes, sexualised narratives or threats of violence to discredit or intimidate them, including through proactive detection tools and partnerships with specialist civil-society organisations.
5. **Prioritise network-level enforcement against divisive AI content farms.** Shift enforcement from a post-by-post approach to network-level action against coordinated operations producing and monetising divisive AI content on

immigration and related issues, including those run from abroad. Meta should invest in detecting cross-page coordination, shared administrators, repeated AI templates and other signals of orchestrated campaigns, and reflect such actions in its transparency reports.

6. **Align AI-labelling and risk mitigation with emerging regulatory benchmarks.** Even outside the EU, Meta should treat the AI Act's deepfake transparency rules and the DSA's systemic-risk framework as guiding standards. This includes: consistent, visible AI labelling; contextual warnings when AI content misrepresents political positions on migration or similar issues; and campaign-sensitive risk assessments that look beyond formal ballots and narrow "crisis" definitions, in line with European guidance on elections and information manipulation.
7.

Referenced research

1. Synthetic Influence – Deepfakes and artificial intelligence in the Hungarian election campaign, Political Capital / Hungarian Digital Media Observatory, April 2026.
2. Circumventing Meta's ban with AI-generated campaign videos and "super forecasting" of Fidesz's defeat, EDMO / Political Capital, March 2026.
3. Normalized Digital Warfare in Hungary was Running Rampant as the Country Approached Election Day, CEDMO, April 2026.
4. Hungary's election is flooded with AI deepfakes – and nobody is stopping them, EU Perspectives, April 2026.
5. Meta removes UK politics deepfake pages after BBC investigation, BBC, March 2026.
6. "How racist AI Facebook posts made one Sri Lankan influencer rich", The Bureau of Investigative Journalism, November 2025.
7. "Patriotic UK anti-immigration social media accounts traced to Sri Lanka and Vietnam", BBC / Panorama, May 2026.
8. Non-Consensual Sexualising Deepfakes – Threats and Recommendations for Legal and Societal Action, CEE Digital Democracy Watch, Policy Paper 2025.

CEE Digital Democracy Watch
Mokotowska 43/104, 00-551 Warsaw, Poland
www.ceeddw.org
KRS 0001090110 | EU Transparency Register 114284692189-05