PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

How the risk of synthetic media that affects global election information is growing

An analysis by PAI

This is PAI's Case Submission to the Synthetic Media Framework. Learn more about the Framework

1 Organizational Background

A contextual introduction to the case study.

PAI'S RESPONSE

Partnership on AI (PAI) is a nonprofit of academic, civil society, industry, and media organizations creating solutions so that AI advances positive outcomes for people and society. PAI is a remote-first organization with over 30 employees distributed across the United States and Canada.

Since its inception in 2019, PAI's AI and Media Integrity (AIMI) program has focused on developing best practices for AI systems that interact with digital media and online information in a manner that serves the public interest, including by empowering the public to interpret synthetic content. We do this by serving as both experts in the field and by catalyzing the community — bringing together key stakeholders from across industry, academia, media, and civil society to produce normative *voluntary* guidance about AI's role in the global information ecosystem. PAI has catalyzed collaboration between sectors on authenticity infrastructure for digital media, studied how audiences interpret synthetic content, and governed technical innovations and research related to synthetic media detection. Further, PAI produces responsive and adaptable governance frameworks and reporting mechanisms for synthetic media creation, development, and distribution.

In February 2023, building upon years of research, convenings, and a focused year-long effort with over 100 institutions across sectors, PAI launched its Responsible Practices for Synthetic Media: A Framework for Collection Action ("the Framework"). The Framework launched with a cohort of ten organizations from across industry, media, and civil society. The Framework identifies three groups involved in the life cycle of synthetic media (**Builders**, **Creators**, and **Distributors**) and provides them with best practices and recommendations on the responsible development and use of synthetic media.

Organizations that have joined the Framework effort have committed to providing a public case study exploring the application of the Framework's principles to a real-world challenge facing industry and society.

The Framework is a living document, and is notably a complement, rather than a substitute, to synthetic media regulation. PAI plans to utilize the input and key takeaways from these case studies to improve future iterations of the Framework to ensure it maintains its applicability and relevance in the fast-moving synthetic media space. Further, we will also consider how to work with industry, civil society, and media to prompt sharing and transparency that moves beyond high-level analysis to detailed, thorny exploration shared with the public.

PAI, as a civil society organization thinking about synthetic media's impact and governance, is contributing its own third-party case study on a pressing challenge for 2024 — the use of synthetic audio in global elections. Our goal in examining three election examples is to apply the Framework's key principles, including those around consent and disclosure, to each and determine where the Framework can help to mitigate potential real-world harm. By doing so, we hope to showcase how the Framework's application to synthetic media in these high-stakes scenarios may support harm mitigation, while also unveiling open, unresolved questions for the field to continue collaborating on.

PAI launched its Responsible Practices for Synthetic Media: A Framework for Collection Action ... with a cohort of ten organizations from across industry, media, and civil society.

2 Challenge

Elaborate on the challenge being addressed in the case study, i.e. the issue to which your organization is applying the Framework.

PAI'S RESPONSE

2024 will be a historic year for global elections, with over 4 billion people, or more than half the global population, eligible to vote in scheduled races. This will be the first large global elections cycle to take place in the midst of the generative AI boom that has led to the democratization of synthetic media tools and capabilities. The barrier to entry for those interested in creating malicious synthetic content is now significantly lower than it was when PAI began its work on deepfakery in 2018; so too, though, is the capacity to create expressive content with AI. As generative AI tools become increasingly accessible and are able to create higher quality synthetic content, their potential for misuse, and even unintended consequences from non-malicious uses in an election context increases.

Here, we examine the use of deepfake audio in an election context in three countries (Slovakia, Pakistan, and the United States) and explore how the Framework can help mitigate real-world harm while exploring issues of disclosure and consent, two key Framework principles. While we acknowledge synthetic media's potential for significant impact on elections (according to one study, 70% of the content sent to fact checkers from three social media platforms during the 2022 Brazilian election was for image or video content, in contrast to text posts, highlighting how prominently media plays a role in elections), we chose to focus on audio due to the high-profile nature of these three examples and the primacy of audio examples in public discourse around the upcoming elections. In addition, while there are technical specificities for the adoption of consent and disclosure interventions for auditory vs. visual content, many of the applications are replicable across mediums.

We first provide the high-level context for each of the three audio deepfakes.

SLOVAKIA

In late September 2023, days before parliamentary elections, an audio deepfake emerged of Michal Šimečka, leader of Progressive Slovakia, in which he allegedly discussed buying votes from Slovakia's Roma minority. The audio, which did not receive Šimečka's consent to be made, was released a few days before the election during Slovakia's 48-hour election silence period. This made the audio difficult for Progressive Slovakia to disprove as it could do no formal campaigning. As described by the ACE Project, an NGO focused on maintaining a repository of electoral knowledge, an election or campaign silence period is "a time frame or a certain number of days immediately before the elections during which no campaigning at all is permitted and there are strict limitations on what the media may write or broadcast." It is unclear what effect, if any, the audio had on the election — however, Progressive Slovakia lost. To date, the creator of the audio is unknown.

PAKISTAN

In early February 2024, as the results of Pakistan's parliamentary elections became public, Tehreek-e-Insaf (PTI) circulated a deepfake audio of its leader, Imran Khan, in which his synthetic voice, superimposed over authentic historical images and footage of him speaking, congratulated his party's supporters for showing up to vote. Imran Khan, however, has been in jail since last year. The deepfake audio *was disclosed as AI-generated with a label. Consent, in this case, is assumed.*

UNITED STATES

In late January 2024, a number of New Hampshire voters received a call with a deepfake audio of President Biden in which he encouraged them not to vote in the democratic primary. The audio, which did not receive President Biden's consent, also made use of one of his catchphrases, "what a bunch of malarkey," in a further attempt to deceive potential voters. Despite not formally being on the ballots, President Biden would go on to win the New Hampshire Democratic primary via a write-in effort. In late February, NBC news first reported that Steve Kramer commissioned the audio deepfake, and that he is a Democratic strategist working on the campaign for Democratic presidential candidate Rep. Dean Phillips. Kramer claimed he commissioned the audio in an act of civil disobedience to raise attention to the dangers of AI in politics.

Many of the potential harms synthetic media can have on elections are listed in Appendix B of the Framework. They include:

- Impersonating an individual to gain unauthorized information or privileges
- **Making unsolicited phone calls, bulk communications, posts, or messages that deceive or harass**
- **Disinformation about an individual, group, or organization**

- Manipulating democratic and political processes, including deceiving a voter into voting for or against a candidate, damaging a candidate's reputation by providing false statements or acts, influencing the outcome of an election via deception, or suppressing voters

(The case study focuses primarily on the harms in **bold**.)

Framework implementation may be helpful (and enforceable by organizations) in cases where known actors are utilizing generative AI systems to produce harmful synthetic content. In instances where *unknown* actors are utilizing *unknown* AI applications, whether open source or proprietary, to produce harmful synthetic content, the incentive for responsible use is minimal. However, the accessibility of open source models and applications built upon those models presents a unique challenge in addressing misuse post deployment, a situation we refer to as the "openness dilemma". This "openness dilemma" remains one of the most important obstacles for those creating and platforming generative AI tools to overcome in order to ensure general responsible use of synthetic content. While closed model and application providers can implement measures to mitigate harmful use after deployment, open source models, once released, can be freely used and modified by anyone, making it more difficult to control their use. We explore how to categorize this challenge by applying the work of PAI's Safety Critical AI team in a later section, and note the need to bridge the synthetic media governance debates with those unfolding related to the availability of AI applications and models.

3 Objective

Describe what your organization is attempting to accomplish by addressing this challenge and/or furthering the opportunities.

PAI'S RESPONSE

These three real-world examples show how the use of synthetic content is already playing a role in global elections. While we cannot determine if election outcomes have been, or will be, swayed by the use of generative AI tools, given their widespread use and potential for real-world harm, their dynamics should not be ignored.

We hope to explore the following questions:

- How can those organizations involved in the life cycle of synthetic media (**Builders**, **Creators**, and **Distributors**) implement Framework best practices to ensure consumers of synthetic media can identify content as synthetic?
- How does consent play a role in each of these examples? Do public figures require different consent protocols for synthetic content?
- What is the broader societal risk to truth and trust that comes from greater ubiquity of synthetic content and awareness that content can be synthesized?
- What interventions can enforce the implementation of best practices for synthetic content transparency, especially by those most likely to cause harm?

As the Framework is a living document, we also aim to identify areas of the Framework that do not provide sufficient guidance in order to update it and ensure its relevance as the uses of generative AI tools continue to grow. Key takeaways will be able to provide those in industry and policy with real, contemporary examples of the use of synthetic content in an elections context and how Framework key principles may help mitigate associated risks.

While it is difficult to determine the role deepfakes played in each of the election examples described earlier, each use case commanded media headlines and highlighted the role synthetic media can potentially play in future elections. That these examples received significant media attention serves as a reminder that news stories can be helpful, complementary disclosure mechanisms to more traditional ones attached to artifacts like labels and watermarks.

4 Framework Scope and Application

Identify which Framework principle was used to help address the challenge/ opportunity, how it was chosen and implemented, and describe how it was applied.

PAI'S RESPONSE

We chose to examine the three election examples while focusing on the Framework principles of disclosure and consent; disclosure services transparency goals, while consent helps support online privacy and the right to one's likeness in the digital age. Notably, how could either of these best practices have been applied to each scenario, and would they have been helpful in mitigating the potential for real-world harm?

SLOVAKIA

Disclosure: Neither direct nor indirect disclosure was applied to the synthetic audio of Šimečka. This, coupled with the release of the audio during Slovakia's election silence period, made the potential for real-world harm or for swaying public opinion very high. Both the creator and the tool they used to make the audio are still unknown, although it was likely a bad actor using an open source model to try and influence election results. This makes accountability all the more challenging. It's also important to keep in mind that bad actors do not have an interest in implementing disclosure practices, so best practices for open source models should be assessed. However, the use of a synthetic media detection tool, such as a classifier (which was ultimately used), by both active and passive **Distributors** of synthetic media (such as social media platforms) as soon as the audio surfaced would have kept in line with the Framework's recommendation on disclosing unattributed synthetic content:

> "Avoid distributing unattributed synthetic media content or reporting on harmful synthetic media created by others without clear labeling and context to ensure that no reasonable viewer or reader could take it to not be synthetic."

Šimečka's team was ultimately able to use an AI speech classifier made by Eleven Labs, an American company that specializes in deepfake audio and text-to-speech generation, to confirm the voice was synthetic. While the use of classifiers may be a useful tool in identifying synthetic media, their use and development can lead to a "deepfake detection dilemma" in which synthetic media detection capabilities become more easily circumventable as they become more accessible.

Consent and Access to a Public Figure's Likeness: As a public figure in a democratic society, Šimečka does not necessarily need to give his consent for the use of his likeness in content that is generally protected by free expression, such as satire and parody. However, the same does not hold true for using his likeness in a malicious context, such as this one, where the intended impact was to disrupt a democratic process. With the understanding that bad actors have no incentive to obtain consent or implement disclosure, the burden of responsibility is then placed on the **Builder** of the (unknown) tool that was used to create this deepfake audio. By implementing content moderation at the point of creation, an emergent best practice, the **Builder** could have helped prevent the creation of the audio. Doing so, however, must weigh the tradeoff of stifling harmful, malicious content featuring public figures while preventing individuals from exercising their right to free expression in creative media that speaks truth to power.

PAKISTAN

Disclosure: In this case, direct, or user-facing, disclosure *was included* in the synthetic content. A visible overlay which said "4th authorized AI voice of Imran Khan" appeared over the video which was distributed by his party. Notably, this was a visual overlay for an audio deepfake. The same direct disclosure mechanism would be inapplicable to a purely audio deepfake which would require an auditory form of direct disclosure. As a best practice, the audio included in this example should have also included an auditory disclosure mechanism as there is a likelihood users may only hear the audio and not watch the video.

IMAGE 1. A screenshot depicting the video of Imran Khan in which his deepfaked voice is played over historical footage of him speaking. Note the direct disclosure in the upper lefthand corner.

While the use of disclosure is a Framework best practice recommended to **Builders**, **Creators**, and **Distributors** of synthetic content, in this case, the video's creator chose to disclose the audio's synthetic nature, as stated in the Framework:

> "Disclose when the media you have created or introduced includes synthetic elements especially when failure to know about synthesis changes the way the content is perceived. Take advantage of any disclosure tools provided by those building technology and infrastructure for synthetic media."

Had the creator not included the overlay, which, admittedly, could have been cropped out or easily hidden, users may have been able to question the veracity of the audio or the fact that Khan was actually jailed and incapable of producing the audio himself. Interestingly, the creator chose to display the overlay in English and not in Urdu, the language Khan is speaking, and it is a visual overlay describing audio that has been synthesized. Ensuring the overlay was also in Urdu would have removed a barrier for non-English speakers to be able to read the disclosure. However, being able to read it does not necessarily translate into being able to understand it — one of the sociotechnical challenges of direct disclosure. This case identifies two possible areas of specificity for the Framework: 1) ensuring direct disclosure mechanisms are provided in the language of the content it is disclosing and 2) recommending auditory disclosure when synthetic audio is played over visuals.

Consent and Access to a Public Figure's Likeness: Consent, in this case, is assumed — the video was distributed by the subject's party and the overlay states the audio is the subject's "*authorized* AI voice." This example raises the question of how consent should apply to public figures when it comes to creation and distribution. Should consent to distribution be implied when it is given for the creation of synthetic content?

UNITED STATES

Disclosure: Similar to the Slovakian example, neither direct or indirect disclosure was utilized in the robocalls using synthetic audio of President Biden. Unlike the Slovakia example, there was no election silence period which meant that Biden's communications staff could, and did, publicly deny the authenticity of the call. The Federal Communications Commission has since banned robocalls that use voices made with generative AI under the penalty of fines, lawsuits, and blocking telephone carriers from facilitating the calls. A company that develops tools to identify synthetic audio claims the audio was made with Eleven Labs' (ironically, the same company that helped identify Šimečka's audio deepfakein Slovakia) technology — highlighting the dual roles and responsibilities held by organizations that sit at the intersection of deeptake detection and tool creation.

Consent and Access to a Public Figure's Likeness: President Biden did not grant his consent to those who used his voice in this case. However, since he is a public figure in a position of power, it is not necessarily required for him to consent to folks using his likeness in content generally protected by free speech. However, his likeness can still be used to cause harm; content moderation can make it such that harm is reduced. Subsequently, Eleven Labs banned the account that was associated with the synthetic audio, but not before the audio was released. Similar to the Slovakia example, this case highlights the need for a form of human moderation at the point of creation, prior to dissemination, ensuring the synthetic audio is generated responsibly. It is worth noting that some companies, such as OpenAI, have content policies and prompt filters in place that prevent users from creating images with politicians — for example you cannot prompt with President Biden. As mentioned earlier, this highlights the need for balancing moderation regimes that take into account immunity from creative expression with the potential risk for harm.

5 Obstacles

Elaborate on any internal or external obstacles intrinsic to the Framework that were overcome.

PAI'S RESPONSE

Disclosure and consent are important tools in the responsible synthetic media toolkit.

Disclosure: Obstacles

Bad actors have no incentive to disclose. Even if they did, there are inherent indirect and direct disclosure obstacles to overcome. For example, each of the indirect disclosure methods highlighted in our Glossary for Synthetic Media Transparency Methods is susceptible to its own technical robustness challenge. The Coalition for Content Provenance and Authenticity (C2PA) standard, while widely adopted, is only useful if implemented by every platform that hosts synthetic content. The moment synthetic content is hosted on a platform that has not adopted the standard, there is less utility for provenance that was added up front. Watermarks can be perturbed or removed, like synthetic audio disclaimers. Further, fingerprinting requires an organization to store the fingerprints created from synthetic content which, given the amount of synthetic content continuously being created, would require massive amounts of data storage.

Indirect and direct disclosure both also face obstacles that are *sociotechnical*. A user hearing synthetic audio may not truly understand what they are listening to without a basic understanding of how the technology works. Further, they may not trust the disclosure they are receiving (see Section 7). How can a voter in New Hampshire tell the difference between a recording of Biden's voice for fundraising purposes and a deepfake audio of Biden asking people not to vote? What types of disclosure mechanisms should they be aware of and look for? Do they understand the limitations of these mechanisms? These questions highlight the need for broader societal education on generative AI tools, their capabilities, limitations, and uses, both responsible and harmful. Trusted learning centers, such as public libraries or universities, can provide a space and resources for the general public to learn more about how these tools are used. And further, models that rely upon community-centric labeling can help alleviate many of the concerns that arise about institutional distrust that reduces the impact AI content labels have on audiences.

Consent and Access to a Public Figure's Likeness: Obstacles

As public figures, politicians' right to consent when it comes to using their likeness in synthetic content is often challenged. Politicians in democratic societies typically have little grounds to require consent for their likeness when used in free speech cases. As stated by WITNESS, "For many democratic societies with a tradition of free speech… Somebody whose words and actions are of legitimate public interest and concern is generally deemed to merit less control over their likeness than an everyday private citizen." However, many providers of generative AI services place prohibitions on allowing politicians to be subjects of synthetic content, largely to protect against potential real world harm. This can lead to a chilling effect on the right of the general public to exercise their free expression.

A possible solution that **Builders** can implement is specialized direct disclosure, such as a unique label that does not interfere with storytelling for synthetic content depicting politicians. This would allow for the use of their likeness in protected mediums but with the added friction of disclosing that the content was synthetic. This would, of course, then require **Creators** to be responsible to consider the potential consequences of their use of synthetic media featuring a politician, and bad actors would likely disregard this best practice.

6 Benefits

Identify the opportunities created for your organization by utilizing the Framework to address the challenge.

PAI'S RESPONSE

The Framework provides organizations with a lens through which to develop responsible synthetic media policies, especially as they relate to elections. Whether an organization builds generative AI tools, creates synthetic content, or distributes/hosts synthetic content, the Framework provides recommendations for how to implement different aspects of responsible use.

The impact of generative AI and synthetic media on elections is difficult to measure and assess. Real-world examples of the malicious use of synthetic media in an election context are already happening, even as policymakers continue to think about how to best address these risks. As generative AI tools become more ubiquitous and easy to use, the likelihood of their use by bad actors will likely increase which, without broadening the general public's awareness of these tools, may lead to long-term societal distrust in the media ecosystem. Exit polling on whether synthetic content has impacted how voters chose to cast their vote may provide some insight into how successful the use of synthetic media in elections can be.

As we examined these three election examples, the "openness dilemma" loomed large, while not invalidating concerns about "closed" models that can be abused. How can we recommend best practice implementation if one of the most common malicious use cases involves bad actors using open source models? However, it is still worthwhile to implement both upstream and downstream mitigations that might prevent some instances of malicious activity. A related question that helps contextualize the impact of disclosure and other best practices which was used as a framing device in one of our PAI meetings was, "why do we lock our car doors even when we know a thief can break in?" Better understanding the impact of open source models, specifically their marginal risk, could support our understanding of the challenges to synthetic media harm mitigation.

To answer the first question we sought guidance from other teams at PAI and looked at how they are approaching similar questions. For example, in our Safety Critical AI team's Guidance for Safe Foundation Model Deployment, the following is described as a baseline recommendation for open source model builders in the pre-deployment stage:

For openly released models, embed safety features directly into model architecture, interfaces, and integrations that cannot be easily removed or bypassed post-release, if applicable for the model's intended domain and task.

By embedding safety features directly into the architecture of a model, such as enhancing a model's technical robustness, open source builders can ensure best practices are baked into their tools making malicious use much more difficult.

The three examples in this case study also highlight the delicate balance needed for enabling synthetic content depicting public figures. Organizations should be able to balance the need for strict policies when it comes to creating synthetic content of public figures that can lead to real world harm while allowing for their use in use cases that are generally protected as free expression. Algorithms that can identify certain restricted key terms are often used but may also be unable to identify the nuance needed when it comes to identifying harmful vs. expressive, protected content.

PAI has identified the following recommendations related to the use of political figures as subject of synthetic content:

- **Builders** of generative AI tools should implement policies that balance access when it comes to political figures — they should allow the usage of political figures' likeness in content that is generally protected in democratic societies with a tradition of free speech such as satire and parody, while restricting that usage in content that can lead to real world harm.
- **Distributors** should not share synthetic content of public officials without at least clearly labeling the content as synthetic. These labels should meet certain requirements in order to ensure their effectiveness and enable user understanding.
- Open source models should follow best practices related to safety pre- deployment, such as red-teaming the model, embedding safety features directly into the models' architectures that are not easily removed or bypassed post-release and, if a frontier or foundation model, staged rollouts to gauge impact before making the model widely available.

The Framework should be able to provide policymakers with a normative framework from which to jump-start drafting and implementing new policies on the responsible use of synthetic media.

7 Conclusion/Key Takeaways

A description of how implementing the Framework ended for your organization, including any lessons learned.

PAI'S RESPONSE

The barrier to synthetic media creation tools for the general public is getting lower while the quality of the content created is only getting better. This means the capability of creating high-quality malicious synthetic content is no longer only in the hands of those with a technical background and financial resources. Without implementing responsible synthetic media best practices such as disclosure and consent, alongside safety recommendations for open source model builders, the risk that synthetic media can lead to real-world harm and/or manipulate democratic and political processes will grow. It is important to keep in mind, as mentioned earlier, that disclosure techniques have their own limitations. For example, according to research from Google's Jigsaw, while labels as a disclosure mechanism may provide helpful context, they may not be the most effective in helping users make decisions about the accuracy of content.

As organizations seek to implement these measures, the risk also grows that users' increased awareness of synthetic content leads people to claim real content to be fake, a concept known as the liar's dividend. In one example from India, a local startup received a request asking the company to create a deepfake using an authentic potentially problematic video of a politician with the goal that the politician could then point to the authentic video and claim it was a deepfake.

Examining these three election examples through the lens of the Framework has helped us identify the following areas for building out the Framework:

- **The importance of content moderation:** Without sufficient safety mechanisms built into the models themselves, human moderation at the point of creation remains the final check on the malicious use of generative AI tools. For example, had Eleven Labs utilized human moderation, it may have caught and prevented the use of its technology to create the deepfake audio of President Biden. While not applicable to open source models which typically rely on built-in pre-deployment safety mechanisms, human moderation can still ensure content that may have bypassed existing security mechanisms is not distributed.
- **Guidance on election silence periods:** Many countries implement election silence periods in the days or hours leading up to an election. Bad actors may take advantage of this period to disseminate malicious synthetic content that attempts to defame or impersonate a candidate, knowing that they will not be able to refute the content. Specific Framework guidance for passive **Distributors**, such as social media platforms, during these periods, on the importance of immediately removing unverified content may provide a necessary level of friction to prevent the rapid spread of misleading media.
- **Consider when it is responsible to use a politician's likeness:** Politicians are public figures. As such, their right to privacy is much more complex than it is for non-public figures. As mentioned in earlier, democratic societies typically consider politicians to have less control over their likeness than private figures. Builders of generative AI tools should develop nuanced policies that account for this distinction when synthetic content featuring politicians is used for protected purposes. These policies should also take into account the potential use of politicians' likeness for harmful and malicious content that is not protected in these same societies such as as incitement to violence, hate speech, and manipulating democratic and political processes.
- **Ensure direct disclosure has intended transparency impact:** The Framework highlights direct disclosure as a key principle. However, it does not further explain how users should interact with disclosure mechanisms or how companies should help users understand them. Overcoming these sociotechnical challenges is just as important as making those mechanisms robust to adversarial and benign risk. Similarly, when it comes to synthetic audio, language matters. In order to ensure disclosure efforts are widely available, all groups identified in the Framework (**Builders**, **Creators**, and **Distributors**) should take steps to ensure disclosures are provided in the same language as synthetic audio and that, per the Framework, aim to disclose in a manner that mitigates speculation about content, strives toward resilience to manipulation or forgery, is accurately applied, and also, when necessary,

communicates uncertainty without furthering speculation. PAI has done previous work on related topics such as how platforms should label manipulated media, an interview and diary study on ecosystem approaches to misinformation interventions, and how these interventions are understood by audiences.

- **The "Openness Dilemma":** Despite the number of recommendations and best practices PAI puts out on the responsible use of synthetic media, bad actors have no incentive to implement them when models, whether open source or proprietary, allow for the creation of malicious content. The Framework currently lacks specific guidance on this front. However, by working internally with PAI teams that do explore best practices and recommendations for model providers, we can offer guidance for open providers that focus on synthetic content on how to best build them with safety and harm reduction in mind. Future work will attend to this dilemma in more depth, considering the unique opportunities and challenges of open foundation models.