PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA

CASE STUDY

How an AI-manipulated video caused harm during South African elections

An analysis by digital democracy nonprofit Code for Africa

CODE for AFRICA

This is Code for Africa's Case Submission as a Supporter of PAI's Synthetic Media Framework.

Learn more about the Framework

1 Organizational Background

1. Provide some background on your organization.

Code for Africa (CfA) is Africa's largest indigenous nonprofit focused on building informed societies. Informed societies need resilient knowledge economies and deeply connected "sensemakers" — people making sense of content and context. CfA helps build these ecosystems by leveraging open data and civic technologies, from AI to drones and other leapfrog innovations, to create enabling technologies and actionable information that contribute to stronger digital democracies.

2 Framing Direct Disclosure at your Organization

1. Please elaborate on how your organization provides direct disclosure (as defined in our Glossary for Synthetic Media Transparency Methods) to users/ audiences.

We do not directly classify as a Builder or Creator of synthetic media. We are a **Distributor** via Code for Africa's iLAB and fact-checking team, PesaCheck. Synthetic media that form part of fact-checking, debunking, or investigative pieces are published on the African Digital Democracy Observatory (ADDO) and Pesacheck websites and shared across social platforms.

CfA's role in identifying potentially misleading or manipulated information includes establishing the origin of the content, how the content was made and edited, and labeling it as such — therefore making use of direct disclosure.

At times, this may include identifying and offering transparency about where content has originated, as well as whether AI tools were used in creating or editing it.

However, we are submitting this case study as an organization that supports informed decision-making by communities, consults with newsrooms, and advises on responsible AI use and declarations of policy to audiences.

Our organization does not directly provide direct disclosure. Our information integrity programs focus on four categories engaged in researching synthetic media that intends to deceive, causes harm, or is partly/fully misleading. In other words, we're more focused on content's meaning and the intent behind it than simply how it has been edited or created. We explore such content in the following ways:

- **PesaCheck** is CfA's fact-checking arm, which uses a series of tools to detect synthetic material that might offer misleading or false information.
- **iLAB** is CfA's forensics investigative and analysis team. It tracks foreign information manipulation and interference (FIMI) networks using social media intelligence (SOCMINT) under the Detection and Information System for Analysing Radicalisation and Misinformation (DISARM) framework.
- **CivicSignal** uses machine learning tools to map and monitor media across the continent.

- **TrustList** offers machine learning methods for reviewing URLs and identifying made-for-advertising sites, which are then flagged with brands for exclusion from programmatic buying.

Using our extensive research into, and analysis of, FIMI and disinformation, we train newsrooms and civil society organizations (CSOs) on how to identify and combat the tactics, techniques, and procedures (TTPs) commonly used by bad actors that use synthetic media in malign ways.

We advise newsrooms and news industry bodies in the following ways:

- The research is used to advise newsrooms on how to develop internal and public-facing synthetic media policies for when synthetic media tools are used in the news production system, or to create content. These policies serve as a user guide for staff involved in the creation and/or publication of such content; as a reference for consumers of the news product; and as a declaration of standards in line with ethical journalism.
- We provide a written internal evaluation process on selecting tools and ethical considerations for using synthetic media. This includes a dedicated section in the organization's editorial policies, covering the newsroom's reasons for adopting AI tools, the ethical and journalistic standards underpinning their usage, and the formats used. We recommend that these policies are public facing and easily accessible by users, as they serve as trust markers to sustain and grow the trust relationship between news organizations and audiences.
- Given the fluid and changing nature of synthetic media, we recommend that newsrooms conduct internal workshops to produce and share policies on synthetic media, and that these be refreshed periodically.
- We recommend that the policy should also include image labeling, with disclosure of what tools are used to create synthetic media. This recommendation is based on the Reuters Institute for Journalism's 2024 Digital News Report, which CfA funds for the four African countries covered. The report shows that concern about what is real and what is fake on the internet regarding online news has risen by 3 percentage points in the last year, with around 6 in 10 (59%) saying they are concerned. The figure is considerably higher in South Africa (81%). Additionally, the top factor globally that respondents said influences their trust in news was transparency regarding how the news is made (71%). The DNR also found that, while audiences tend to be uncomfortable with the use of AI to create new content, not all forms of content are seen equally. The report found that people were least resistant toward the use of AI to generate text-based content but most strongly oppose the use of AI for creating realistic-looking photographs — and especially video, even if disclosed. For CfA, this means that direct disclosure plays an important part in allaying mistrust in the media.
- Our recommendation to organizations that provide direct disclosure is for their

policies to explain how they use and label synthetic media. We recommend that they identify all usage of synthetic content on the assumption that nondisclosure, even of benign usage rather than malicious usage, has the potential to erode the trust relationship between organizations and their audiences.

- This labeling should not be intrusive, and should follow established media standards for crediting media. But the information should be easily available to users who wish to establish the provenance of an image, video, or audio artifact. This follows closely what we understand from the C2PA recommendations and the PAI Framework.

Further below, we provide a real-world example of an instance in which potentially misleading content was not directly disclosed in a responsible manner — and its impact.

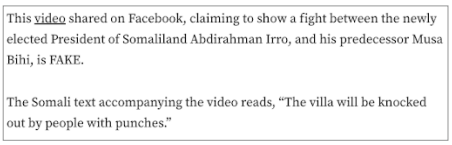
---

2. Does your organization understand the goal of direct disclosure as specified in the PAI Framework: "to mitigate speculation about content, support resilience to manipulation or forgery, be accurately applied, and communicate uncertainty without furthering speculation" or does it have a different understanding?

CfA implements direct disclosure as a **Distributor** of debunked synthetic media that was created with the intent to deceive or misinform. We do this by adding a label above the image on the published article. When it's published as a result of fact-checking, the headline and article blurb include information that the image has been manipulated.

We also recommend to newsroom and CSO partners that they adopt a public-facing policy on direct disclosure and implement it according to that policy. This aims to provide mechanisms for legitimate actors/organizations to preemptively build resilience, rather than attempting to restore trust retroactively.

Headline and blurb of a debunked video on PesaCheck



This video shared on Facebook, claiming to show a fight between the newly elected President of Somaliland Abdirahman Irro, and his predecessor Musa Bihi, is FAKE.

The Somali text accompanying the video reads, "The villa will be knocked out by people with punches."

Body of text also debunking the claim
Source

3. What, if anything, from your organization's approach to direct disclosure is missing from this NIST taxonomy below? Should it be added to a taxonomy of direct disclosure? If so, why?

From NIST's Reducing Risks Posed by Synthetic Content:

> The most commonly used techniques to *directly disclose* to the audience how AI was used in the content creation process include:
>
> - content labels (e.g., visual tags within content, warning labels, pre-roll or interstitial labels in video and/or audio, and typographical signals in text highlighting generated AI text with different fonts),
> - visible watermarks (e.g., icons covering content indicating AI usage where the bigger the icon, the harder its removal), and
> - disclosure fields (e.g., disclaimers and warning statements to indicate the role of AI in developing the content, and acknowledgments to provide more context to the AI contribution and credits to reviewers).

CfA would recommend going a step further in the disclosure fields for news media: declaring and explaining their use of synthetic media in their editorial policies/newsroom values.

4. What criteria does your organization use to determine whether content is disclosed? What practices do you follow to identify such content?

CfA recommends that all synthetic content be disclosed to avoid potentially eroding audience trust in news media. We recommend that this be done with clear captioning at minimum, according to the news publication's established system of disclosure and crediting for all content. Newsrooms should consider adopting "layers" of identification — C2PA-based, for example — on a preestablished scale of editorial ethics.

Established media standards provide a useful starting point. For instance, consider tonal corrections and cropping as acceptable. In the UK, The Guardian describes it as "anything that may have been done in a darkroom."

CfA, iLAB, and PesaCheck use a range of tools to identify content that has not been disclosed as a way to debunk disinformation or to understand tactics used by bad actors to inject synthetic media into an information sphere.

These items are debunked and published on their sites with captions indicating that they are synthetic or have been manipulated.

5. Per the Framework, PAI recommends disclosing "visual, auditory, or multimodal content that has been generated or modified (commonly via artificial intelligence). Such outputs are often highly realistic, would not be identifiable as synthetic to the average person, and may simulate artifacts, persons, or events." How does your organization's approach align with, or diverge from, this recommendation?

Given our focus on newsrooms and CSOs, we recommend full disclosure of all content created or edited with generative AI. This can be as simple as adding a caption to the image in the standard publishing format, but it must always be backed up by a public-facing editorial policy explaining the organization's ethical and technical parameters for such usage. This would include material that is intended to mislead for the purpose of satire.

The biggest difference in PAI's recommendation for disclosing and our recommendation is the context in which content created or edited with generative AI is being shared. As we are concerned with the news context, we believe that all content should be labeled, regardless of whether it is unrealistic or realistic, designed to mislead or not. Again, this process need not be intrusive. Fundamentally, this is a structure designed to inoculate news organizations against accusations of manipulating information, and to give their audiences the comfort of knowing that there are fixed, accessible editorial standards and policies governing the news organizations' content production.

3 Real World, Complex Direct Disclosure Example

1. Provide an example in which your organization applied (or did not apply) a direct disclosure to a piece, or category, of content for which it was challenging to evaluate whether it warranted a disclosure (based on your organization's policy). This could be because the threshold for disclosing was uncertain, the impact of such content was debatable, understanding of how it was manipulated was unclear, etc. Be sure to explain *why* it was challenging.

In May 2024, South Africa's then-largest opposition party, the Democratic Alliance, broadcast its final advertisement ahead of the general elections. The video was shared on all its social platforms and broadcast on the various platforms of the public broadcaster, South African Broadcasting Corporation (SABC). The advertisement shows what The Mail & Guardian referred to as "a computer generated version of the South African flag burning."

The ad received major criticism from citizens and a range of CSOs, decrying the "unpatriotic" act of burning the flag. While burning the flag is not illegal in South Africa, the outcry led to an announcement from the public broadcaster that it would not air the ad again. The president went as far as to call it "treasonous."

Screenshot from YouTube

The decision whether to label or not label this as synthetic media should have been considered in a larger context. The general election was the first in South Africa involving fear and concern over the use of synthetic media to manipulate voters or influence the outcome. To date, the Democratic Alliance does not seem to have considered that labeling the flag-burning section of the video as synthetic media would have mitigated the pushback from citizens and the state, including that from the public broadcaster. As this is an advertisement rather than editorial content, the public broadcaster should not be held responsible for affixing a label in order to maintain editorial independence. The responsibility should lie with the creator.

CSOs held numerous conversations with the Independent Electoral Commission (IEC) to safeguard the integrity of the elections and hold political parties accountable to do the same. As a show of good faith, this political party (the second largest in the country) should have labeled this video of the burning flag as synthetic. The flood of news coverage about the burning flag removed the weight of the messaging and led to government resources being spent on investigations into potential repercussions. These were political repercussions focused on retribution for the perceived denigration and desecration of national symbols, rather than discussions on the video content and the nature of synthetic media.

For instance, the public broadcaster SABC refused to air the ad. The Independent Communications Authority of South Africa (ICASA)'s complaints and compliance committee later found that the SABC's refusal to air the ad had "no legal basis" under the Electronic Communications Act or the regulations dealing with political advertising, and recommended that the SABC be fined ZAR 500,000 (USD 27,000) for banning the ad from its public service television channels due to "prejudice caused to the DA."

2. How was this piece/kind of content identified?

During general elections across the countries we operate in, we monitor social and other platforms for potentially manipulated information. This includes monitoring of political party messaging and advertisements, any electoral regulatory bodies, and the incumbent government. In cases like these, a small team of analysts do regular scans of official social accounts — dependent on national election processes — for narratives and messaging. Media reporting is also followed by CivicSignal MediaCloud to see how the messaging is reported. The impact of the messaging guides whether further investigation is required.

CfA, at the time, did not further investigate this campaign, but rather monitored the media coverage surrounding it as there was no clear attempt to spread harmful manipulated media.

3. Was there any potential for reputational (e.g., negative impact on your organization's brand, products, etc.), societal (e.g., negative impact on the economy, etc.), or any other kind of harm from such content?

The advertisement had the potential to cause both reputational and societal harm. The political party faced heavy criticism for the imagery of the burning flag, which might have been less pronounced if it had been declared that no actual flag had been burned. The advertisement also gave ammunition to actors invested in promoting false or misleading narratives inimical to democratic processes. This extended to Zuko Madikane, a human rights lawyer, asking the Gauteng High Court to declare the DA in gross violation of sections of the Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000, saying that its message "seeks to propagate violence and incite harm to society."

Expanding further on Section 3, Question 1: In the months leading up to the elections and directly afterward, the Electoral Commission of South Africa (IEC) faced increased attacks from bad actors claiming that it was compromised, which helped make the elections very contentious. For example, ex-President Jacob Zuma's uMkhonto Wesizwe (MK) party refused to accept the outcome of the May 29th national election, claiming vote-rigging, and said that more than 9 million votes were unaccounted for in the election. The party launched an interdict against the IEC, alleging substantial election rigging during the recent national elections.

4. What was the impact of implementing this disclosure? How did you assess such impact (studying users, via the press, civil society, community reactions, etc.)? Did the disclosure mechanism mitigate the harm described in the previous question (3.3)?

This is an example of nondisclosure. If there had been direct disclosure, we believe that the Democratic Alliance — and, more important, its supporters on social media — could have defused many of the attacks by pointing out that the video was clearly marked as synthetic media, with no actual flag burned. We are not suggesting that this would have reduced the volume of attacks; one may assume that there would still be outrage at the metaphorical burning of the flag. Importantly, however, this would have been a way for the DA to empower its supporters to respond from a position of shared trust. We are dealing with a political party here, but the trust mechanism would hold for the broader relationship between news organizations and audiences. It is difficult to assess the impact that direct disclosure would have had, given that no comparative example relates to this specific situation.

5. Is there anything your organization believes either the Builder, Creator, or Distributor of the content should have done differently to support direct disclosure?

Our example is one of nondisclosure. However, the creator of the content would have benefitted by using a tool that allowed for labeling or otherwise signifying that this was synthetic media. Potentially, this was an oversight, and the backlash was unexpected. In South Africa, the Advertising Regulatory Board has a Code of Conduct that Creators and Distributors are governed by.

South Africa's Department of Communications and Digital Technologies has published a national policy framework for AI, on which public submissions are being taken. This process is moving slowly, so adjusting the Code of Conduct for Creators and Distributors would come as one step in guiding advertising agencies on how to disclose synthetic media.

6. In retrospect, would your organization have done anything differently? Why or why not?

In the example in question, the Democratic Alliance should have directly disclosed, both in the caption of the video and in a banner overlay on the actual video, that parts of the video were synthetically created.

The video was created and formatted for a number of platforms in addition to TV such as Facebook, X, and YouTube. These social platforms already have a self-reporting mechanism to label synthetic media, which the creators could have used. For TV broadcasts, a simple label running at the bottom of the ad would have been sufficient.

7. Were there any other policy instruments your organization relied on in deciding whether to, and how, to disclose this content? What external policy may have been helpful to supplement your internal policies?

Since our fact-checking organization is part of the International Fact-Checking Network (IFCN), we use their guidelines to build our internal policies.

After joining PAI, we also used much of the Synthetic Media Framework for policies — especially in our advisory role to newsrooms when we researched the prevalence of editorial policies in newsrooms across the continent.

Some of our policies are also informed by the DISARM Red Framework's approach and the Stanford Institute for Human-Centered Artificial Intelligence, at which CfA has regularly participated in events.

8. What might other industry practitioners or policymakers learn from this example? How might this case inform best practices for direct disclosure across those Building, Creating, and/or Distributing synthetic media?

This example makes a case for the disclosure of synthetic media so as to preempt attacks on content credibility, avoid confusing audiences, and to maintain a relationship of trust with audiences. The lesson of this example, for media practitioners, is that disclosing the use of synthetic media alters the dynamic between producers and consumers of content. In this case study example, there was no conscious attempt to fool viewers into believing that the flag burning was real. But disclosing that it was synthetically produced would have allowed the Democratic Alliance to refute accusations of having physically destroyed a South African flag. This is just one example of how total transparency about the use of synthetic media can affect the trust relationship.

This is further discussed in Section 3, Questions 3 and 4.

4 How Organizations Understand Direct Disclosure

1. What research and/or analysis has contributed to your organization's understanding of direct disclosure (both internal and external)?

In 2022, CfA conducted a survey across three African countries (Kenya, South Africa, and Zambia) to ascertain news media editorial policies. The methodology was twofold: We interviewed newsroom managers and conducted desktop research of 30 to 40 newsrooms per country. The study was intended to gauge media transparency to promote trust in the newsroom, as well as to assess vulnerabilities to influence and attacks on media freedom.

From our research, we discovered that fewer than 20 percent of the sampled newsrooms in the three countries fully published their editorial policies. The majority of newsrooms had published parts of their policies, but further interviews indicated that newsroom staff had not been trained to use them.

Out of roughly 120 newsrooms sampled, none had an internal AI usage policy or an external-facing policy dictating to their audience how it might get used, how it would be declared, and what personal data it might collect. However, from the interviews, more than half of the newsroom managers indicated that their journalists used AI-driven tools (such as Grammarly). This research was conducted before the mainstream release of ChatGPT and comparable tools.

An updated version of this study will be done in 2025.

2. Does your organization believe there are any risks associated with either OVER or UNDER disclosing synthetic media to audiences? How does your organization navigate these tensions?

We believe that underdisclosing synthetic media creates significant risk for organizations that rely on a trusting relationship with users. There is potential for an unfortunate corollary in that overdisclosing means that synthetic media is foregrounded, thereby allowing malign actors to benefit from calling into question the authenticity of organic media. Underdisclosure, however, would have a much more fundamental effect on trust, with specific reference to news organizations. As the Reuters report cited in Section 5, Question 2 below says, if we don't provide audiences with information they may want so as to help them decide what news to use and trust, this will be at least equally damaging as overdisclosing the information.

Identifying synthetic media created with the intent to deceive would require a different sort of labeling than for synthetic media that is an editorially sanctioned part of an organization's production cycle. For synthetic media intended to deceive, the labeling serves as a warning and/or disclaimer. For synthetic media that is editorially created and curated, the labeling serves as a disclosure of editorial policy.

3. What conditions or evidence would prompt your organization to re-calibrate your answer to the previous question (4.2)? E.g., in an election year with high stakes events, your organization may be more comfortable *over labeling*.

Given that our example is one from a political party's messaging as part of an election cycle, and that our key takeaway from this example is that direct disclosure would have brought benefits for the party and its supporters in terms of counter-messaging against attacks on social media, we are proponents of over labeling. However, we feel that a system to label synthetic media needs to be robust enough to weather a sudden surge of interest in high-stakes events, rather than changing to fit circumstances. If the system is not fit for purpose when it is under stress during such events as an election, it won't be fit for purpose at all. Again, we are specifically commenting from within the news ecosystem, where news organizations need to ensure that the trust relationship with audiences is as secure as possible.

4. In the March 2024 guidance from the PAI Synthetic Media Framework's first round of cases, PAI wrote of an emergent best practice: "Creative uses of synthetic media should be labeled, because they might unintentionally cause harm; however, labeling approaches for creative content should be different, and even more mindfully pursued, than those for purely information-rich content."

   Does your organization agree? If so, how do you think creative content should be labeled? What is your organization's understanding of "mindfully pursued"? If your organization does not agree, why not?

In general, we feel that context matters. Creative content framed in the context of entertainment or satire could be negatively impacted by direct disclosure. But because our emphasis is on news organizations and CSOs that rely on a fragile relationship of trust with their audiences, we suggest that for news organizations, a policy of labeling everything at a basic level (e.g., a caption on a photograph for synthetic media that is outside the agreed standards discussed in Question 5) is desirable in that context.

An argument can be made that standards might differ when synthetic media is used for creative or satirical purposes, but we still recommend a minimum disclosure — crucially one that is clearly mandated by the publication's own editorial policy. Ideally, news organization policies should map to the policies of synthetic media creators, such as advertising agencies. The information arena encompasses more than those two stakeholders, of course, and strategic collaboration on this by all stakeholders could be a way to make this happen.

5. Overall, what role(s) does your organization believe Builders, Creators, and Distributors play in directly disclosing AI-generated or AI-edited media to users?

We feel that this should be a shared responsibility, mapped to agreed-upon, mutual policies and standards. In the instance of our case study, an agreed-upon standard for labeling synthetic media for Builders, Creators and Distributors, based on an imperative to maintain audience/consumer trust would ensure a three-stage process making all parts of the chain of production equally responsible. We recognize that the PAI Framework lays out and encourages this. This would simply provide the mechanism for labeling synthetic media according to an agreed-upon methodology.

6. How important is it for those Building, Creating, and/or Distributing synthetic media to all align collectively, or within stakeholder categories, on a singular threshold for:
    1) the types of media that warrant direct disclosure, and/or
    2) more specifically, a shared visual language or mechanism for such disclosure?

    Elaborate on which values or principles should inform such alignment, if applicable.

This is vital. Without a shared standard for disclosure, media consumers are left uncertain. The effectiveness of media literacy programs is also reduced. And as with such anti-disinformation methodologies as the DISARM framework, a shared understanding of synthetic media is needed in order to effectively engage with, and combat, synthetic media that is intended to cause harm.

Organizations such as Partnership on AI, C2PA, Stanford HAI, etc., all contribute toward a larger understanding of "shared language."

5 Approaches to Direct Disclosure, in Policy and Practice

1. What does your organization believe are the most significant socio-technical challenges to successfully achieving the purpose of directly disclosing content at scale? (Refer to question 2.3 for reference to PAI's description of direct disclosure)

One of the primary purposes of implementing direct disclosure is to mitigate erosion of the trust relationship between news media and audiences. But measuring the efficacy of direct disclosure is difficult. Does it affect trust negatively or positively? What impact does it have on how the content is valued? Another challenge is how to accurately and usefully measure trust in media and the impact of that trust relationship on the business of news. Labeling synthetic media, as the Reuters Institute has pointed out, is "an instance where it is relevant and important to disclose it. But that can be challenging because we don't really know how audiences are going to respond. Transparency can sometimes be a difficult thing to navigate."

2. What is your organization hoping to accomplish by implementing direct disclosure? Does your organization believe directly disclosing ALL AI-edited or generated media, is useful in helping accomplish those goals?

We are proponents of direct disclosure because of the threat that nondisclosure poses to the robustness of information environments. As the Reuters Institute for Journalism puts it: "Carefully threading the needle when it comes to disclosing the use of AI will be crucial for publishers concerned with audience trust, as will be explaining to audiences what AI use in journalism looks like. Excessive or vague labeling may scare off individuals with already low trust and/or those with limited knowledge about what these uses entail, who will likely default to negative assumptions. **But failing to provide audiences with information they may want to decide what news to use and trust could equally prove damaging.**" *(Emphasis added.)*

3. Please share your organization's insight into how direct disclosure can impact:

   1) Accuracy
   2) Trustworthiness
   3) Authenticity
   4) Harm mitigation
   5) Informed decision-making

   Note: You can also discuss your understanding of the relationship between these concepts (for example, authenticity could impact trustworthiness, harm mitigation, etc.)

For us, the primary consideration is how direct disclosure (or lack of direct disclosure) can impact trustworthiness. Authenticity and accuracy are a subset of this in that direct disclosure can bolster the qualities of authenticity and accuracy, which in turn help build trust. For us, authenticity and accuracy both enable building a relationship of trust with audiences — and the currency of trust. Our benchmark for the relationship between direct disclosure and authenticity and accuracy is included in the section above. To summarize: Failing to provide audiences with information they may want in order to decide what news to use and trust could damage the trust relationship.

Direct disclosure also aids in mitigating harm, both for the audience and content creators. This is not just in the obvious sense of enabling audiences to avoid harm related to disinformation, but also by decreasing opportunities for bad actors to weaponize synthetic media.

4. Does your organization believe there will be a tipping point to the liar's dividend (that people doubt the authenticity of real content because of the plausibility that it's AI-generated or AI-modified)? Why or why not? If yes, have we already reached it? How might we know if we have reached it?

The answer to this depends on several factors. First is context: Is the information environment being manipulated to drive doubt regarding authenticity, and what is the sociopolitical framing for that? Second comes incentives and the notion of fungible truth: Are people incentivized to introduce doubt into the information ecosystem (e.g., for the purposes of political grandstanding) to affect shifts in the sociopolitical ecosystem? There is no such thing as a tipping point for this, merely a flux. In other words, bad actors will saturate the information system with misinformation and/or disinformation — or take advantage of existing saturation — when they can. Whether they succeed will depend on how the information ecosystem has built its checks and balances, and how it has made this process available to citizens of the ecosystem.

5. As AI-generated media becomes more ubiquitous, what are some of the other important questions audiences should be asking in addition to "is this content AI-generated or AI-modified," especially as more and more content today has some AI-modification?

Crucially, audiences should be asking, "*Why* has this piece of content been AI generated." If we believe that, for example, AI-produced journalism should have the ethical standards and constraints that journalism in general does, then the consumption of AI-generated content should be subject to the same norms of critical consumption.

6. How can research help inform development of direct disclosure that supports user/audience needs? Please list out key open areas of research related to direct disclosure that, the answers to which, would support your organization's policy and practice development for direct disclosure.

We need more practical research — especially focused on specific global regions — that can quantify the impact on the trust relationship between audiences and creators of synthetic media when that media doesn't carry direct disclosure. This means measuring whether audience trust diminishes over time when Creators do not disclose, and whether it increases for Creators that do disclose.

6 Media Literacy and Education

1. In the March 2024 guidance from the Synthetic Media Framework's first round of cases, PAI wrote of an emergent best practice: "Broader public education on synthetic media is required for any of the artifact-level interventions, like labels, to be effective."

   Does your organization agree? If so, why? Has your organization been working on "broader public education on synthetic media"? How? (please provide examples.) If your organization does not agree, why not? What responsibility do organizations like yours (identified in the Framework as either a Builder, Creator, or Distributor) have in educating users? What about civil society organizations?

We are not sanguine about the effectiveness of media literacy education, synthetic or otherwise. In our view, expecting citizens to shoulder the burden of identifying misinformation or synthetic media, for example, is a form of victim blaming. We believe that the creation and dissemination of synthetic media should be subject to restrictions according to best practice and agreed upon by relevant industry bodies such as editors guilds. The consumption of synthetic media will then entail learning by doing, rather than actively educating. This is not to say that media literacy programs are not worth promoting, just that our primary focus should be on the production of a safer, structured information environment.

2. What would you like to see from other institutions related to improving public understanding of synthetic media? Which stakeholder groups have the largest role to play in educating the public (e.g., civic institutions, technology platforms, schools)? Why?

The largest stakeholder in educating the public is the public itself. We feel that a robust methodology for creating and disseminating synthetic media (created by, and from the point of view of, news organizations, which are our primary concern) will provide the public with the tools to construct its own understanding. In other words, citizens who comprehend the role synthetic media plays in a news environment will be able to correct those who are willfully or accidentally promoting the manipulation of information.

3. What support does your organization need in order to advance synthetic media literacy and public education on evaluating media?

The establishment of an agreed set of rules and standards for production and disclosure will be key to advancing literacy.

7 Commentary on the Framework's First Set of Cases (beyond Direct Disclosure)

1. The first round of cases did not just focus on direct disclosure, but also on broad exploration of several case themes: creative vs. malicious use, transparency via direct and indirect disclosure, and consent. We want to leave room for respondents to highlight any other areas of the Framework that can be deepened or improved upon to ensure its viability in a rapidly changing synthetic media ecosystem (related to the case themes above, and moving beyond the direct disclosure focus of this case template).	N/A
2. Has putting the Framework into practice influenced other processes, procedures, or policies at your organization?	We use the PAI Framework as part of our resources for newsrooms that are building their policies. We often cite the previous round of case studies — especially the CBC News case study — as ways to incorporate thinking about synthetic media from the perspective of journalistic values.