PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA

CASE STUDY

How an investigation in South Asia uncovered harmful synthetic media

An analysis by technology nonprofit Meedan

Meedan

This is Meedan'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.

Meedan, a 2024 Skoll awardee, works to create a more equitable internet. We do this by developing innovative open-source collaboration technology and by implementing creative, multistakeholder projects with network civil society actors, researchers, policy stakeholders, and technologists. We deploy these programs during critical civic and social moments, including elections, public health crises, and natural disasters, to ensure that communities can access and evaluate the information needed to make informed decisions. Our 42-member team, working across 14 time zones, has led some of the largest collaborative journalism projects in the world — enabling our work across 45 countries and in 31 languages. Meedan builds technical tools to support the workflows of journalists and fact-checkers, who may also be considered Distributors or Builders of synthetic media as defined by PAI's Responsible Practice for Synthetic Media Framework. Meedan is submitting this case as a civil society Framework supporter that also builds technology, within which our users may engage in the disclosure of AI-generated content.

Our work with AI

At Meedan, we're leveraging generative AI (GenAI) in unique ways. Many people associate large language models (LLMs) with activities such as text or image generation. However, we primarily use GenAI in the realm of multi-class and multi-language classification tasks; we've found LLMs to be invaluable tools for this work. The ability of these models to follow intricate instructions has allowed us to build prototypes that effectively classify content submitted through tiplines and sort content into user-defined taxonomies.

The use cases we serve demonstrate the versatility of GenAI beyond its more publicized creative applications, and they showcase the technology's potential to change the way we organize and enrich information for our end users. Based on user-defined taxonomies and using limited human supervision, ClassyCat — our internal classification and categorization tool — leverages LLMs to generate labels for social media content. (See our Special Interest Group on Information Retrieval 2024 demo paper for more details.)

However, our efforts with GenAI classification aren't without challenges. We're actively working to address issues related to language complexity, classification specificity, and cultural context. For instance, can ChatGPT, Claude, and others process texts in Arabic or Bengali as skillfully as they do in English? What is the best way to incorporate knowledge into classification taxonomies? How can our system deal with a lack of specificity? Are we mitigating the potential harms, biases, and cultural insensitivities of LLMs?

These hurdles underscore the importance of developing AI systems that can navigate the nuances of diverse languages. Using LLMs in this way can reduce barriers to entry and allow people with less technical expertise to create classifiers. As we continue to refine our approach, we're committed to creating more sophisticated, context-aware AI systems that can take into account the complexities of global communication and information flows. Ultimately, we're interested in AI applications that help journalists do their work more efficiently.

Spotlight on Check

Over the past decade, Meedan has been building Check, an open-source platform for journalists, civil society organizations, intergovernmental groups, and other stakeholders to connect with audiences on closed messaging apps like WhatsApp. Check helps partners better understand information patterns by taking in and providing an interface for examining user-submitted content. It also supports their processes for responding to users with algorithmically driven, human-in-the-loop workflows. One of the key pillars of our work is our support for newsrooms, fact-checkers, and civil society organizations around the world as they collect, annotate, and assess misinformation and disinformation. Check supports these information stakeholders through the deployment of tiplines on messaging apps. Organizations use these tools to collect requests for fact-checking from their audiences and then aggregate and analyze submissions, as well as respond to them — all from within Check.

Newsrooms and community organizations around the world now use Check for digital verification, investigations, information distribution, citizen science workflows, monitoring in moments of civic importance, and other large-scale journalism projects, as well as to share information with communities affected by crises.

One of the key functions of Check is that we provide infrastructure for organizations to create their own white-labeled tiplines and to distribute newsletters. Such infrastructure is also being used to better understand the information needs of communities and as tools for citizen science. They can collect examples of content that community members witness in their daily lives and help users uncover information patterns related to key topics such as gendered disinformation, hate speech, misinformation and disinformation, and other harmful content.

> *The use cases we serve demonstrate the versatility of GenAI beyond its more publicized creative applications, and they showcase the technology's potential to change the way we organize and enrich information for our end users.*

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.

Meedan's role involves providing the tooling (Check) that newsrooms and fact-checkers use to intake, examine, and respond to queries, questions, and requests for verification of misinformation and disinformation — all of which can include synthetic content. Our journalism partners have identified that in fact-checking, what the content alleges and who created it are often more important than whether or not it's synthetic. Rather than being the primary goal of a journalistic output, discerning whether a piece of content contains synthetic elements is one layer of context that would be incorporated into verifications or assessments that a fact-checker might create.

With regard to our partners and users, as well as our data and media literacy training approach, Meedan's disclosure policy is centered on transparency. Internally, we disclose any use of automation in our product. Within Check — which is focused on text-based item annotations, content clustering, and the act of matching queries with responses — our user interface has a history log that identifies what actions were taken by which stakeholders for a particular item in the organization's workspace. We disclose any use of AI as part of the workflows that Check users may experience. AI similarity-matching technology and manual-matching features that allow users to cluster similar content are embedded in Check. These functions can be helpful for assessing the prevalence of submissions to tiplines that contain variants of a specific piece of synthetic media. We disclose AI-suggested matches to our partners. We also have visual cues that mark when automated tagging has been used for matching an item to a particular cluster or category. To show that this has taken place, we use the lightning bolt emoji (⚡).

When our partners identify, through their own internal verification workflows, the presence of GenAI content among user submissions, they can use item statuses and human-authored reports to disclose that the content they're responding to was synthetically generated. For such disclosures, we do not use overlays. Instead, the journalist's assessment is, in essence, the disclosure, and we create individual links that bring the audio file, image, or video together with that assessment.

As the topic areas that we prioritize and work on sometimes include subjects that involve providing critical and life-saving information to end users, we want to make sure that our tooling focuses on streamlining information-distribution processes for Check users. Our goal is for Check users to be able to distribute and disseminate information to their communities as efficiently and effectively as possible. As such, we leverage AI to support back-end workflows, such as categorizing content; to date, Meedan does not use AI to generate content that will be shown to the general public.

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?

Broadly, in the case of **implementing** generative technology, Meedan’s approach is to do no harm. This approach involves the absence of interventions or additions to an information-distribution workflow that may cause adverse outcomes.

When it comes to the **disclosure** of generative content, we employ a harm-reduction approach, opting for strategies and practices that aim to reduce the negative consequences of a given intervention. This translates into risk reduction for creating unintentionally false or misleading content, as well as the promotion of transparency to the fullest extent possible between creators and their audiences.

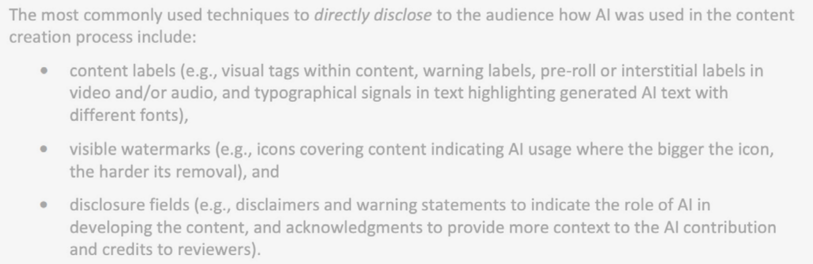
In pursuit of harm reduction, our understanding of the goals of direct disclosure is aligned with what is outlined in the PAI Framework. The accurate application component is of greatest importance to us. In the future, if we want our tooling to support the generation of new answers from existing content, we will be fully committed to making sure those responses are accurate.

In terms of Meedan’s own work, at present, we use GenAI only to help create text-based classifiers for different types of content, such as misinformation. As we do so, we aim to be as transparent with Check users as possible to identify where AI has been utilized as a component of a verification or question-and-response workflow. Although the generative technology that we use internally influences processes, rather than being visible to users — and it is a text-based technology, as opposed to using other forms of media — this same approach still applies.

As one example, in the field of collaborative verification, it is incredibly important to track who has done what. In a similar manner, Check logs and displays the history of an item in its user interface. This history incorporates information about who has engaged in the verification workflow of an item that is set to be fact-checked and what steps that person has taken. The log clearly shows when our internal AI has acted on an item, as opposed to a human (i.e., when a topical tag has been added to an item in Check to cluster it with other, similar items in a given journalist’s or fact-checker’s workspace). We also use visual cues in our interface to identify when these internal systems have supported a human fact-checking workflow. For example, tags added by AI are prefaced with a lightning bolt emoji (⚡), whereas tags added by a human are not. The lightning bolt was originally selected to indicate the speediness of auto-tagging.

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).

We believe it is important to include the following elements of context as components of disclosure fields:

- **Attribution for any tools or software** used in assessing the provenance of synthetic media.
- **Attribution to any human investigative conclusions** used in the assessment by citing or linking to source image provenance, or through the use of other techniques.
- **Training data transparency,** which is key to media literacy in the short term. In the long term, it will be essential to the socialization of current and future generations into an information environment where human-made content is used to generate synthetic content, and whether, by default, we should have an expectation of knowing what data informed the creation of any synthetic content we encounter.

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

For our own internal technology, we do not have any plans to engage in the creation of any tooling that is specifically tasked with the detection of synthetic media, but we have partners who investigate and assess synthetic media. At times, they may leverage third-party tools to make these assessments; these are details that they keep internal as part of their own verification workflows. As our partners, they would be making a disclosure that would be related to such assessments, and we encourage them to include details on their processes in those written assessments, including any third-party tools they may have used. We do not currently provide recommendations to partners on technologies they should implement as part of their workflows, but we would be interested in such recommendations from the PAI community.

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?

We are supportive of disclosures based on the description above. However, because many of our partners are journalistic organizations doing reporting, fact-checking, and assessment on submitted content, disclosures are secondary to the journalistic output. Disclosures may be, in some cases, components of such outputs (e.g., articles, explainers, fact-checks, or resources). In situations wherein users submit synthetic media for fact-checking, the issue of an image's provenance may be a central aspect of the journalist's output and would be expected to be supported with evidence (e.g., links to the source image, evidence of alteration, etc). In general, we are very supportive of disclosure when an audio file, image, or video reflects deceptive or harmful intent, whether the content targets an individual or is intended to operate at a societal or civic level.

3 Real World, Complex Direct Disclosure Example

1. Provide a real-world example in which either: a) Direct disclosure should have been applied, or b) Direct disclosure was applied to a piece, or category, of content for which it was challenging to evaluate whether it warranted a disclosure. 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 is challenging.

As an organization, we do not engage in the development or identification of generated media. The examples below were collected as part of a research project that Meedan contributed to on gendered disinformation in South Asia, which was conducted in collaboration with Chambal Media, The Quint, and Digital Rights Foundation.

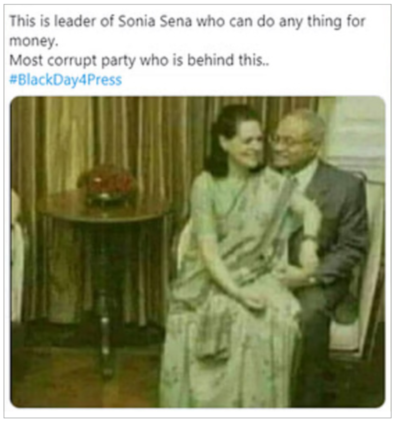
Internally in Check, annotations from partner organizations indicated that these examples contained synthetic components. The first image (Example 1) was published in 2020, prior to the advent of disclosure implementation by social media platforms. The second image (detailed in Example 2) was published in 2024 and was not labeled or disclosed as synthetic on social media.

The purpose of this investigation was to examine existing barriers to the detection of harmful content focused on gendered disinformation, and to both propose and carry out a new methodology for developing definitions for harmful content that can be operationalized. While synthetic media detection was not the primary goal, some examples of gendered disinformation that were discovered, archived, and annotated by the team of researchers did contain manipulated or synthetic features that were not disclosed as such on social media.

In annotating a database of examples, project researchers hypothesize that more effective, robust, and informed definitions (in this case, based on the topic area and lived expertise of practitioners in South Asia working on reporting and responses to gendered disinformation) of what constitutes different forms of harmful content that can be integrated into a taxonomy. They can also be used in computational assessments to support better algorithmic detection and labeling.

Both examples involve reputational harm for both the individuals and their professions, as well as societal harm — potentially discouraging equitable political and journalistic participation for women and minority gender identities.

Example 1: A manipulated image to tarnish the reputation of a political leader (published in 2020).



This is leader of Sonia Sena who can do any thing for money.
Most corrupt party who is behind this..
#BlackDay4Press

- Description of the image: Indian political leader Sonia Gandhi is pictured sitting on the lap of the former Maldivian president with the claim that she will do anything for money.
- The claim associated with this image: “This is leader of Sonia Sena who can do anything for money. Most corrupt party who is behind this. #BlackDay4Press.”
- Context: This post is using an image that has been photo-edited to show Sonia Gandhi sitting on the lap of the former president of the Maldives. In the original photo, they are sitting apart in separate chairs. This image was not clearly labeled as having been manipulated.

Example 2: An AI-generated image to assassinate the character of a journalist and her source (published in 2024). *Project partners, to date, are not sharing this image to prevent further publicizing the gendered disinformation.*

- Description of the image: An AI-generated image claimed to have been “leaked” portrays a journalist in Pakistan wearing revealing clothing (organizations working on the annotation requested that this journalist remain anonymous) ahead of the 2024 elections in the country.
- The claim associated with the image: “[Journalist name] election se ek hafta pehly ki photo leak ho gayi.” (Translation: “[Journalist name] photo leaked a week before the elections.”)
- Context: A photo circulates with the claim that it is a leaked image of a journalist wearing revealing clothing ahead of the 2024 elections in the country. While the image is left for interpretation, the implication is that she is exchanging sex for access to information related to her reporting. This image was not labeled/disclosed as a generated image.

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

Supported by the Sexual Violence Research Initiative, Meedan led a study to examine gendered disinformation in South Asia and create a database of such instances. This regional context required the creation of new, operationalizable definitions for gendered disinformation related to the topic. Our research has demonstrated the importance of hyperlocal definitions and community context in order to modify the inclusion criteria for harmful and hateful content in ways that are more relevant to the affected communities. As part of this project, collaborators used Check to document instances of gendered disinformation in South Asia.

Contributing partners collected posts organically through their own social media feeds, as well as through submissions from community members and as they encountered artifacts in their daily work. The latter are collected by running tiplines, conducting fact-checking for reporting, and soliciting content from networks of reporters, rural networks with access to grassroots communities, and activists working at the intersection of gender and technology. All three organizations were led by women at the time of the project, and all self-identified as feminist groups while supporting different audiences and networks across India and Pakistan.

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?

Clear types of harm specific to this sort of content are outlined in the research links shared below (Section 3, Question 4), both for the immediate target and for society at large. As will be outlined in our forthcoming research paper on this investigation, such harm ranges from reputational damage to the incitement of violence against the individuals who appear in the images, as well as toward others who might be implicated in a broader harmful narrative. It may result in more widespread and societal harm related to public participation. Further research is required to understand whether, for these specific examples, disclosure of the synthetic elements of the images would have mitigated harm.

Our collaborative investigation highlighted gaps in how existing operationalizable definitions used to detect harmful content are developed, and offered a new methodology for improving them.

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)?

In collaboration with our local partners, we developed an annotation schema in which team members go through a deliberative evaluation process to ascertain whether any particular piece of content is actually an example of gendered disinformation. To do this, they examine the claim, target, harm, and source. In this case, whether or not something was AI-generated was identified with a tag that partners would add to a specific piece of content. Identification was based on meeting criteria that were developed by contributing partners in order to capture the relevant context of South Asia:

Nature: Gendered disinformation is a manifestation of online gender-based violence that can also include offline activities.

Modus operandi: Relies on and promotes misogynistic, sexualized, false, and deceptive narratives influenced by a variety of intersectional social identities such as caste, religion, sect, and gender.

For this effort, whether or not the artifact was synthetic was less important than what it was depicting.

The focus of this work involved assessment and annotation of the content for the purpose of categorizing subthemes within the concept of gendered disinformation. A deeper examination of context, which includes the impact of gendered disinformation among communities, was conducted by each contributing organization and published in the form of three case studies. These case studies reference additional examples to those outlined in Section 3, Question 1:

- Chambal Media: "Disinformation and Disempowerment: The Gendered Experience in Rural India"
- The Quint: "Harassed, Yet Resilient: Muslim Women Journalists Fight Gendered Disinformation"
- Digital Rights Foundation: "Gendered Disinformation in South Asia case study - Pakistan"

Recommendations from this project identified key areas where social media platforms could improve their monitoring, screening, and community contributions, namely through escalation channels where hyperlocal insights in diverse contexts and regions can be captured to ensure that harmful synthetic content receives adequate disclosure such as labels.

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

In Meta's 2024 memo on practices for the disclosure of generative AI, it mentions that a public survey revealed "support for labeling AI-generated content and strong support for a more prominent label in high-risk scenarios." We agree with this approach, and we argue that prominent, high-risk disclosure across contexts is essential. From Meedan's operational flow, determining whether or not a piece of content represents a high level of risk is an assessment undertaken by community partners that conduct fact-checks, distribute resources, or create other journalistic outputs and that best understand the hyperlocal contexts, languages, and nuances of the regions they serve.

For certain topics, we recommend a unified, consistent approach for disclosure and labeling. Meedan's Digital Health Lab project worked with different search and social media platforms to identify health-related topics of importance. Current algorithmic and human-supported moderation and labeling efforts prioritize only topics that make headlines. Although underrepresented topics with significant health impacts may not be high priorities, they may sometimes pose a high level of risk. As such, for topic areas where the risks presented by misleading content may be higher overall, we recommend consistent and prominent labeling that discloses whether content was algorithmically generated. This includes gendered disinformation instances that may result in negative implications for free and fair elections or lead to an incitement of violence.

In addition to engaging in disclosure practices, we recommend the implementation of clear media literacy interventions that address norms around false positives and negatives. Given the current standards and limitations of today's available research, we are not able to reliably determine whether content is authentic or not, and detection mechanisms will never be 100 percent accurate. As such, models can have both false positives and false negatives. False negatives and unlabeled synthetic content may appear to be more believable to users and ultimately be more deceptive as a result. The types of detailed disclosure outlined below could support a better understanding of where content has originated, which could help with this problem. That said, since overly sensitive moderation may generate false positives, clearly available mechanisms should be accessible so posters of authentic media can appeal this label. Ideally, creators generating the content would be able to disclose whether or not their creation leveraged AI. Because this is an evolving ecosystem with such a wide variety of stakeholders — potentially including bad actors — the nature of decentralized creation makes this process difficult.

For content requiring disclosure, we recommend that Builders of models used to create synthetic content identify the fact that data was used in its generation and say what data was used. Adding a "Where did this come from?" disclosure feature to generated content is suggested. (This can be related to a variety of data points: the geographic location inspiring the piece and additional data that informed its creation.) A key area of importance and priority for us is to reference where the data that was used to develop a particular piece of content came from. In this way, we hope to make sure that consumers of media, including synthetic media, are socialized in an environment wherein stakeholders have a more nuanced understanding of this concept than simply attributing the media's origin to broad, undefined "artificial intelligence processes."

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

In addition to stakeholders who are tasked with disclosure supporting the development and creation of more equitable detection systems, Meedan recommends supplementing existing labeling and disclosure actions with infrastructure both to appeal labels and to escalate requests for labeling. Suitable processes can be used to appeal whether the content has been generated or augmented by AI, which would add a new and important layer to existing content-appeals processes for misinformation and disinformation and for hateful or harmful speech. It is important to note that Meedan, in collaboration with PEN America, has examined existing reporting mechanisms on social media platforms in the past and has found significant challenges and barriers to effective escalation pathways.

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?

In this context, fact-checkers and human rights defenders working to address disinformation relied on internal workflow processes for annotation of harmful content, as well as disclosure. No additional policy instruments were necessary for their work, as identification and disclosure within the annotation process was a component of verification.

For the social media platforms, PAI's Framework should serve as a guide for disclosure. If there were best practices for community inclusion in detection efforts or specific requirements for engaging in community level consultation, such documents would be a helpful addition to the field.

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?

Meedan feels strongly about the importance of creating inclusive spaces where Larger World (Meedan's preferred terminology for "Global South") perspectives — and other hyper-local or context-specific experiences — are able to be shared. In a similar way to the process outlined above regarding our research initiative on gendered disinformation, a component of more equitable disclosure should involve more engagement in processes for creating detection algorithms, which can be used to detect harmful content in general. This can look like the joint creation of operationalizable definitions for high-risk topics for detection and disclosure prioritized by community organizations that represent affected groups.

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)?

With existing Meedan partners and Check users, we have referenced PAI's Framework and covered disclosure policies distributed by social media platforms. We are currently engaging in a GenAI learning exercise with our network of community partners to understand their experiences with generated content, their use of detection tooling, and their processes for verifying and disclosing whether a piece of content has been generated. This will help advance ongoing conversations about GenAI usage and current practices in the fields of journalism and fact-checking response.

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?

Over disclosure and under disclosure, mislabeling, false negatives, and false positives all pose a risk to perceptions of trust in our online information ecosystems and could contribute to distorted understanding and misrepresentations of reality. Our concern is that these issues will disproportionately affect communities that are already underserved or underrepresented in technology development and infrastructure. This is where many members of Meedan's network of community and civil society organizations come in. Incorporating topic-area and lived expertise into the development of models that train detection systems may lead to more inclusive, representative, and accurate training data. We believe strongly that improving disclosure systems requires greater civil society engagement, including involving affected communities and their representatives in the process of informing harm and risk taxonomies. As a second component of this work, Meedan researchers and

PEN America have identified the limited effectiveness of existing appeals and escalation channels on social media platforms. This is an avenue we are actively researching with our partners in order to build better mechanisms for reporting and addressing issues with labeling or automated detection.

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*.

The volume of content that is likely synthetic, as well as improvements in the certainty of labeling synthetic content, could shift this balance. There's a need for empirical research on how labeling affects the perception of unlabeled content and the overall trust in content online. Ultimately, there are trade-offs in various labeling approaches in terms of balancing false positives and false negatives. The "best" choice is unclear, and may well vary between audiences and topical domains. We believe empirical research is the best approach to understanding these trade-offs.

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?

"Creative" is a complex term. Creators can be creative in both benevolent and malicious ways. In the same way that information-rich content can be used to target, harass, or harm individuals and communities, creative work can be used for these purposes. Meedan sees no difference in how content should be labeled, regardless of whether the content was for creative use. If there is a high risk of harm, regardless of whether the content in question might be deemed creative or informational in nature, it should be subject to consistent labeling practices. Content generated and distributed under a properly labeled parody or satire account can be easily transported out of that platform and distributed to audiences who take the content at face value.

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?

Our response in Section 3, Question 5 outlines the additional practices that Meedan recommends.

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.

It's critical that stakeholders align their practices for sharing insights based on user research. Visual standards benefit end-users and take the guesswork out of understanding divergent visual signals. This alignment, however, assumes a uniformity of surfaces and communities, so it may not be possible in practice. Policy and community-representative gatherings that aim to bring together relevant stakeholders can help to create the necessary conditions for alignment. Short of regulations that support alignment or adherence to a framework, community coordination is essential. The creation of collaborative datasets — modeled on structures such as the Global Internet Forum to Counter Terrorism (GIFCT) — represents one possible element of supportive alignment infrastructure as well. The benefit of easing pressure toward alignment is that by implementing and expressing differing responses and interventions, we can gain a wider understanding of how end-users and the public will respond to generated content.

5 Approaches to Direct Disclosure, in Policy and Practice

1. What does your organization believe are the most significant sociotechnical 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)

Meedan believes that direct disclosure in a manner that, per the PAI Framework, strives to "mitigate speculation about content, support resilience to manipulation or forgery, be accurately applied, and communicate uncertainty without furthering speculation" is an important priority. Essential challenges here are that disclosure labels:

- Are being accurately applied (mitigating false positives and false negatives, as outlined in Section 3, Question 5).
- Include opportunities for appeal. Existing appeals processes and processes for reporting have been identified — including by our own internal researchers — as limited, but this will be an important mitigation for over labeling.
- Are accompanied by risk mitigation and media literacy plans to address mislabeling. Inevitable failures gain outsized media attention and undermine the trust end-users have in disclosures; both false positives and false negatives can be extremely viral media events. Preparation for such events is essential across stakeholders.

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?

Internally, Meedan discloses the use of AI to Check users directly within our tools, which are currently used to support internal workflows for human rights defenders and journalists. We are committed to transparency and disclosure as we develop more GenAI-powered features such as text descriptions of images, keyword/phrase extraction, claim extraction, and video summarization. In all of these initial cases, the generated content will be visible only to fact-checkers in our web application, where we can probably label it for these experts.

Few partner organizations using Check use GenAI to create content sent to their end-users on WhatsApp and other platforms. We are currently engaged in a learning process with some of our network partners to understand how they engage with, and verify, whether content has been generated, the tooling they use for this process, and their disclosure practices for end-users.

We do reference key documentation, including PAI's Glossary for Synthetic Media Transparency Methods, and highlight documents describing responsible practices. Ideally, this documentation could be localized into some of the languages that our partner network members prioritize.

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.)

Here, we will continue with insights derived from the previously outlined disclosure example about gendered disinformation in the Larger World, and we'll reference some of the challenges associated with underrepresented or underserved contexts and languages.

Authenticity: This is a precursor to many of the other concepts, which have as the goal or foundation the ability to better detect and validate authentic content.

Accuracy and trustworthiness: Here, as mentioned, the challenges associated with false positives and false negatives — and the varying perspectives users hold about labeling infrastructure — will inform successful implementation. These patterns may differ across contexts, so it will be important to examine implementation strategies from a social norms perspective in a variety of languages and locations, as well as on diverse platforms. Labeling accuracy, which includes labeling whether something is or is not synthetic, will also vary depending on the training data. We recommend collaborating with organizations that have the local experience — and the hyperlocal datasets — to uncover whether certain types of content may pose a high degree of risk in one locale or context but not another. Who owns these datasets, and what can be done with them, are secondary considerations that we are working on closely with our network of partners. Moving forward, they would be very interesting questions to explore as a PAI community. The insights derived from such datasets may have implications for what recommendations related to disclosure implementation would look like.

Harm mitigation: In order to address the challenge of scale, Meedan's collaborators in the fact-checking and journalism fields are increasingly seeking support for how they can respond to overarching narratives rather than just individual pieces of misinformation. Often the content's message is as important as the generation process. Most of our partners focus on what is communicated and factual errors in the statements made, rather than whether the content is synthetic. That said, being able to demonstrate that content is synthetic can help debunk false content in some cases. Focusing on the more persistent narratives allows our partners to offer context on a claim regardless of the medium. For example, the gendered disinformation in our South Asia project identified various generated images communicating the same idea: that women in politics are compromising their bodily integrity to achieve political goals, perpetuating a larger narrative that undermines women's legitimacy in civic and social participation. In such cases, researchers who contributed to this project asked whether this broader narrative contains more "harm" than the individual claims or generated images addressed or disclosed in isolation. Are there particular types of harm that, when combined, are more problematic than the sum of their parts? Is it useful to disclose whether a given generated image is part of a larger narrative? How should this context be communicated? Should these be prioritized by review for moderators or flaggers? Meedan is actively engaging with risk typologies that align with the Framework, especially when designed in a way that pulls from public health prioritization and harm-reduction approaches.

Informed decision-making: As an organization, we believe that an equitable information ecosystem provides users with access and context to support informed decision-making. This involves combining labeling with other media and data literacy components.

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?

This is an empirical question that should be researched further. In particular, we are concerned that an abundance of synthetic, false, or misleading content could lead to general distrust and disengagement with online content. We would like to see experimental approaches analyze this question and examine the factors at individual and content levels that can contribute to this. Broadly, given the limitations of today's available research, we are not able to reliably determine if content is authentic or not, and detection mechanisms will never be 100 percent accurate. This makes the trust and authenticity challenges particularly salient. Research that examines the relative behavioral impacts of excessive false negative or false positive content can help us understand which combined computational and literacy-related interventions can help mitigate these impacts.

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?

Many of Meedan’s partners are currently working in the fields of journalism, communications, and fact-checking. In fact-checking, it’s often more important what the content alleges, and who created it, than whether or not it’s synthetic. That is, very harmful content can be generated by either humans or AI, as can trivial, non-harmful content. For our partner organizations, it’s more important to assess the message of the content and its potential harm. (See the examples in Section 3 for more detail.)

The main way that detecting and labeling AI-generated content helps is by encouraging audiences to be skeptical of content that is generated by AI. In general, however, we want people to have a healthy dose of skepticism for all content and to develop good media literacy skills.

We are also particularly interested in the broader norms that internet users are being socialized into as large training datasets are used to develop AI-generated content. Meedan supports a community-driven data ownership model whereby local partners steward and sustain collective knowledge ecosystems without the exploitation of commercial AI. We are actively working with our partner organizations to develop creative ways for generating knowledge datasets that can be managed, owned, and used by our partner organizations in various ways across civil society and technology. Disclosure that adds depth to a piece of content — identifying both the fact that the content was generated and also indicating what was used to generate it — is a first step in the important component of media literacy.

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.

Key open areas of research that would support our understanding of direct disclosure include:

- Exploring how community-informed datasets can be operationalized so that more diverse users can be supported through automated disclosure mechanisms.
- Identifying which content truly poses a high degree of risk in terms of informing norms, decisions, and behaviors.
- Examining how patterns may differ across platforms, and uncovering the expectations that different users have with each platform.

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?

Meedan agrees with this. During webinars for feature rollouts, we have been building awareness about these topics among our partner organizations. We are also in the process of building a chatbot specifically designed to support crisis preparedness with some of our community partners. This chatbot leverages content created by our partner network. It aims to provide information to broader publics across various languages about topics such as how to archive and document social media content, as well as how to identify and protect oneself and one's communities against misinformation and disinformation, including from misleading synthetic content.

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?

We would like to see more education and transparency. Openness about the datasets used to generate synthetic content will be foundational as we work to set expectations and establish appropriate social norms around the use and development of synthetic media. Prototyping different ways of offering this transparency would be a valuable initial exercise in discovering what actually helps people understand this process in a digestible way. Meedan is particularly interested in how this conversation intersects with discussions about data ownership, especially as we think about the training data integrated into the detection algorithms used to identify synthetic content.

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

Being able to contribute their firsthand experiences to policy dialogues around synthetic media and media evaluation would be a great knowledge-sharing and capacity-building opportunity for our network of partners, especially those who work on the evaluation of misinformation and disinformation or the aggregation and analysis of harmful speech. If there are any opportunities to engage in this way, it would be great to have these direct experiences inform policy recommendations or to provide our partner network with concrete ways that they can contribute to advancing this topic.

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).

We are excited about the opportunity to engage in conversations related to more community-informed detection, text-based disclosures, and the data sources informing generative AI.

2. Has putting the Framework into practice influenced other processes, procedures, or policies at your organization?

We are working with our network of partners toward the creation of programming related to generative AI, disclosure, and awareness of the challenges and trade-offs in disclosure policies across the platforms they use for communication, verification, and fact-checking.