PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

Even the best-intentioned uses of generative AI still need transparency

An analysis by human rights organization WITNESS

WITNESS
SEE IT FILM IT CHANGE IT

This is WITNESS' Case Submission as a Supporter of PAI's Synthetic Media Framework.
Learn more about the Framework

1 Organizational Background

A contextual introduction to the case study.

WITNESS'S RESPONSE

WITNESS is an international human rights organization that helps people use video and technology to protect and defend their rights. Our Technology Threats and Opportunities Team engages early on with emerging technologies that have the potential to enhance or undermine society's trust in audiovisual content. Since 2018, WITNESS has led a global effort, Prepare, Don't Panic, to understand how deepfake and synthetic media technologies and, more recently, large language models (LLMs) and generative AI, are impacting at-risk communities around the globe, and how responding to real world harms and risks can bolster human rights. These efforts have included contributions to the development of technical standards, pioneering work facilitating real-time analysis of suspected deepfakes that can have important consequences for democracy and human rights, input to technology companies' policies and legislative proposals, experimentation with generative AI tools for human rights advocacy, public advocacy, and in-depth consultations with activists, journalists, content creators, technologists and other members of civil society.

As a civil society supporter of Partnership on AI's (PAI) Synthetic Media Framework that does not directly create, host, or distribute synthetic media or its underlying technologies, WITNESS is submitting an analysis of a third-party case example. WITNESS has no affiliation with the creator of the case study selected, and it was in no way involved with the development of the content, its publication, its distribution, or its moderation. The case example was selected because of the complex issues that it raises regarding the use of synthetic media and its pertinence to the ongoing development of PAI's Framework.

This case was the focus of a workshop discussion organized by WITNESS in Bogotá, Colombia, in August 2023. The analysis provided here reflects this discussion.

2 Challenge

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

WITNESS'S RESPONSE

Over the past five years that WITNESS has led the Prepare, Don't Panic initiative, we have been exploring the opportunities that synthetic media can create for the communities we center and support. These opportunities need to be weighed against potential negative effects, including those affecting human rights, so that we may advance ethical and responsible behavior through mechanisms such as PAI's Framework.

The selected case study pertains to AI-generated images created by the IAbuelas account on Instagram. Between 1974 and 1983, the military junta in Argentina hid or killed 30,000 people of all ages and social conditions. Among them were hundreds of pregnant women who gave birth in captivity, as well as girls and boys who were kidnapped along with their mothers and/or fathers. Around 500 children born in detention or kidnapped along with parents were delivered to families close to the Armed Forces or security, or abandoned at institutions. As the parents of the children had been killed or disappeared, the grandmothers took it upon themselves to find their grandchildren, thereby creating the organization "Abuelas de Plaza de Mayo" (Grandmothers of the Plaza de Mayo). The association created a public database with photographs of the dissidents who had been killed and whose children had been abducted. Drawing from this public database, artist Santiago Barros created AI-generated images of what the missing children might look like today and published them on Instagram.

Barros, whose own family was affected by the dictatorship, created the images using Midjourney, allowing him to combine the images of parents to generate a synthetic portrait of their missing child as an adult. For those situations in which the gender of the baby was not registered, the artist generated an example for both male and female genders. Barros' stated aim has been to "stir the conscience" of those aged over 46 years and older who may have doubts about their own background, as well as to underscore the decades of work done by Abuelas de Plaza de Mayo.

Barros posts the images regularly on the account IAbuelas on Instagram, which is now a private profile at the request of Abuelas de Plaza de Mayo. To date, 255 posts have been published, though their status as AI-generated is not clearly disclosed in the images. The image captions, while not explicit about Barros' use of AI, do include several hashtags that allude to their synthetic origin, such as #ia (the Spanish translation of the acronym for AI) and #midjourney.

The organization Abuelas de Plaza de Mayo has praised this initiative, but they have also warned that the only infallible tool to connect the missing grandchildren with their families of origin is the DNA matching that continues to be carried out by the National Genetic Data Bank (Banco Nacional de Datos Genéticos).

This case first came to our attention when it was picked up internationally in mid-2023 by Rest of World. In the following days the case was covered by multiple outlets, including El País, AP, and France 24.

3 Objective

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

WITNESS'S RESPONSE

IAbuelas represents a case of how synthetic media and generative AI tools can be used for human rights advocacy. Its status as an artistic project aligns with the current criteria for responsible use of generative AI as outlined in Appendix A of the PAI Framework. Although no harm from the IAbuelas case has been publicly reported, unintentional harm is a serious possibility arising from artistic projects that lack prior consent and/or fail to clearly communicate their synthetic nature to audiences.

Recognizing the laudable aims of Barros' project, the case highlights important questions, notably those around the unintended consequences from the use of AI. Of particular importance, in our view, are the implications of cases such as IAbuelas for two themes in PAI's Framework: consent and disclosure.

With regard to consent, there are two primary issues we wish to examine: (i) responsible use of generative AI tools when they are employed to edit or generate content on the basis of personally identifiable information taken from public databases; and (ii) consent when data relates to an individual who is deceased or kidnapped. It is worth noting that, in the IAbuelas case, potential concerns around licensing are not applicable, given the data was public. This case brings questions around the value of seeking consent even when data is publicly available, as a means of averting potential harm. It also puts under examination the role that external consultation might play in responsible and ethical practices for imagining futures, specifically when victims' of human rights abuses are the subject of the creation.

In terms of disclosure, our main concern was whether or not the **Creator**, the social media platform, or the **Builder** of the AI technology had done enough to effectively signal that the images were generated with AI and that they did not reflect the actual appearance of the abducted children. A lack of Direct (user-facing) Disclosure can enable nefarious uses of synthetic media and lead to unintended harm. In this case, a lack of clearer labeling as well as, possibly, a lack of Indirect Disclosure (not user-facing) could have sown confusion about the actual appearance of the abducted children, thereby inhibiting efforts to locate them.

Considering the above challenges, we seek to encourage responsible ways of creatively leveraging synthetic media to raise awareness of human rights issues by:

1. Exploring guidelines on consent that mitigate and avert the potential for harm when using online public archives.
2. Analyzing different forms of disclosure to further develop best practices and expectations across a pipeline of actors.

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.

WITNESS'S RESPONSE

CONSENT

The artist accessed the online archive of the Abuelas de la Plaza de Mayo, an organization founded by the grandparents of the abducted children, to find and synthetically merge photographs of their parents in order to generate images of how the children would look today.

In this case, consent can be considered from the lens of the Abuelas de la Plaza de Mayo, whose online archive was used; the parents, whose photographs are merged to create images of their children; and the missing children themselves, whose identity is being referred to in the images published by the artist. According to various media reports, the artist did not obtain consent from the Abuelas de Plaza de Mayo organization or any other relatives before initiating the project or publishing it online.

The current provisions in PAI's Framework on consent do provide a solid basis to address this case. For content creators, it includes the need to be transparent about how they obtained consent, except for "reasonable artistic, satirical, or expressive ends." While we believe that the PAI Framework balances a concern for freedom of expression with the need to address potential harms, the current draft could be strengthened by emphasizing the benefit of seeking consent when the likeness of real people is directly involved in the input or output of the AI generation process. This should not be mandatory, and there can be circumstances in which consent may not be pertinent, feasible, or even needed. WITNESS has offered guidance on informed consent in the context of human rights documentation and advocacy.

Another issue raised by the IAbuelas case, one not currently addressed by the PAI Framework, is how consent should be handled if the person in question is deceased or missing. We believe that in these cases it is necessary to take intentional steps to respect the individual and what their preferences might have been. Although there is no clear-cut way to know the preferences of the deceased or missing, contacting relatives, a person's estate, or next-of-kin could be a proactive step in that direction. This approach has been adopted in prior situations, for example by Propuesta Cívica, when they constructed a deepfake of murdered journalist Javier Váldez.

Lastly, we believe that the current provisions for **Builders** of AI Technology and Infrastructure, as well as for **Distributors** and publishers, are enough to address concerns around consent exemplified in this case.

DISCLOSURE

Although active distribution channels (i.e., the media outlets that carried the story) were sufficiently clear about the origin of the images, PAI's Framework does provide guidelines that could have aided the content creator, the passive distribution channels (i.e., Instagram), and the **Builder** of the AI technology to be more transparent about the source and history of the images. The only label for viewers about the origin of these images on Instagram were the hashtags "#ia" and "#midjourney" in the caption under the images.

Current provisions in the Framework to disclose, or facilitate disclosure, include but are not limited to:

- 5.0 [For **Builders** of technology and infrastructure]: Take steps to provide disclosure mechanisms for those creating and distributing synthetic media.
- 11.0 [For **Creators**]: 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.
- 12.0 [For **Distributors** and publishers]: Disclose when you confidently detect third-party/user-generated synthetic content.

In addition, the PAI Framework clarifies that disclosure can be direct (user-facing) and/or indirect (not user facing), and it offers specific examples that can be facilitated or included by the different actors across the pipeline. The content creator could, for instance, add a visible watermark directly to the images to communicate their creation with AI. At WITNESS, we have been thinking about creative ways of using visible forms of disclosure to strengthen artistic expression. One example to consider is the case of *Welcome to Chechnya*, a documentary where the creators added a halo effect around the face of people whose identity was protected with face-swapping technology.

As it is, PAI's Framework offers enough provisions to argue that the content **Creator**, the passive **distribution** channel, and the **Builder** of the technology did not do enough to disclose to viewers the origin of the images. However, the range of direct and indirect mechanisms listed, and the limitations and nuances involved in each of them, may prove overwhelming and unclear for the different stakeholders involved. This is especially so for content **Creators** who, as in this case, may have less resources or experience to comply.

This ambiguity is arguably a reflection of our understanding of these technologies when the PAI Framework was first published. The rapid pace of development in this field means that now, only a few months later, we may be able to consider more specific guidelines for responsible and ethical use. In addition, since releasing the Framework, PAI has released further guidance on indirect disclosure methods and options.

Lastly, given that existing direct and indirect disclosure mechanisms exhibit certain limitations requiring further research, development, and experimentation, we posit a "toolbox" approach to synthetic media disclosure and detection — i.e., for **Builders** of technology, for **Distributors** and publishers, and for **Creators**, to be encouraged to enable and/or use more than one disclosure mechanism to offset shortcomings. In the case of IAbuelas, for example, Instagram could have added a visible or audio label on the image for viewers on their platform, and a fingerprint in case the image was shared externally, as occurred with subsequent media coverage.

We posit a "toolbox" approach to synthetic media disclosure and detection — i.e., for Builders of technology, for Distributors and publishers, and for Creators, to be encouraged to enable and/or use more than one disclosure mechanism to offset shortcomings.

5 Obstacles

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

WITNESS'S RESPONSE

There are various obstacles that have been discussed throughout the process of creating the PAI Framework that are worth reiterating: synthetic media and its underlying technologies are still in an early, yet rapidly evolving stage; the mechanisms to address harms are even less developed and refined; and we have yet to see the scope of use cases and gray areas that can define this space and, therefore, expectations for responsible use in regulation and similar AI governance products. The similarly evolving nature of this PAI Framework already reflects a recognition of these obstacles.

It may seem that the lack of enforcement mechanisms and flexible wording can render PAI's Framework ineffectual. However, we believe that openly recognizing that it is not a substitute, but rather a complement, for developing legislation, internal policies, and other enforceable mechanisms that promote accountability and protect human rights is an important element to continue highlighting in future versions.

6 Benefits

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

WITNESS'S RESPONSE

The PAI Framework provided us with a set of consensus-driven recommendations that can be applied at any stage of the lifecycle of synthetic media. As external actors, we are able to bring to bear the PAI Framework's operationalization with **Builders**, **Creators**, and/or **Distributors**. It has given us a tool to utilize in our discussions with organizations seeking to develop synthetic media responsibly. For example, by working with startups seeking to build generative AI tools with consent in mind, we can point to the consent recommendations in the PAI Framework as a useful starting point. As long as it continues to be updated and maintains its relevance (with input from a global and diverse range of stakeholders), we will be able to utilize the PAI Framework in our engagement activities with industry actors and governments.

7 Conclusion/Key Takeaways

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

WITNESS'S RESPONSE

CONSENT

The PAI Framework includes provisions that address most of the concerns around consent that emerged from analyzing the case of IAbuelas. The following are areas of improvement that we have identified:

- Although making exemptions for artistic, creative, and expressive ends is a necessary provision to ensure freedom of expression, obtaining and communicating consent should still be recommended to avert and mitigate unintentional harm. Highlighting exceptions, such as in the case of political satire, would be relevant. WITNESS offers guidance on informed consent and how to obtain it
- Guidance on consent when dealing with the likeness of a deceased or missing person can help address gray-area cases.

DISCLOSURE

The current provisions on disclosure offer a general overview of the existing possibilities but they do not set clear expectations. Although this reflects the stage where we are in terms of the development of these technologies, we suggest making the following changes:

- Propose that **Builders** of technology, **Distributors** and publishers, and **Creators** should enable and/or use more than one disclosure mechanism to offset shortcomings.
- Include a provision to highlight the need to develop standardized and interoperable solutions.

GENERAL

We believe it is essential to recognize that PAI's Framework does not replace the need to develop legislation, internal policies, and other enforcement mechanisms that promote accountability and protect human rights. Highlighting this understanding can strengthen future versions of this document and help ensure the PAI Framework is a complement to these efforts.