PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

How TikTok launched new AI labeling policies to prevent misleading content and empower responsible creation

TikTok

This is TikTok's 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.

TIKTOK'S RESPONSE

TikTok is a dynamic entertainment platform empowering 1B+ users worldwide to express themselves through content, connection, and creativity across our diverse global community. This includes enabling creators to upload AI-generated content (AIGC), made off-platform, as well as developing effects or features that may use AI to provide a vibrant and engaging creative experience. Therefore, per Partnership on AI's (PAI) Synthetic Media Framework, we would fall under the **Builder** and **Distributor** categories.

Our objective in developing our user-facing policies (Community Guidelines) around synthetic/AIGC is to balance creative expression with harm mitigation, and to find an appropriate way for creators to disclose their use of AI tools. As a platform that both hosts AIGC and makes AI tools for creators, we want to create a culture around synthetic media that empowers creators to explore positive uses of AI, while encouraging transparency via self-disclosure, and making clear what use cases of this technology are harmful. We already prioritize transparency, from "sponsored" content labels and unsubstantiated content labels, to "state-affiliated media" labels, and verified badges that inform our users of the authenticity of notable figures' accounts. When it comes to AIGC, transparency is particularly important to help prevent users from being misled. This is especially nuanced when it comes to cases like political satire — specifically content that, when labeled, is clearly humorous, but if not labeled, could easily mislead viewers about the truth of events.

2 Challenge

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

TIKTOK'S RESPONSE

One significant challenge was anticipating potential harms presented by AI tools as well as the positive, or non-harmful, potential uses of synthetic media. TikTok's Integrity and Authenticity Policy team developed the first iteration of our synthetic media policy in the autumn of 2022, primarily in an anticipatory fashion; at the time, there was increasing awareness of synthetic media creation tools, but they were not widely available and AI content had not yet appeared in high volumes on online platforms. While we already prohibited material that had been edited in a way that might mislead users about real world events, we believed there was potential for harm (both for individuals depicted in synthetic content related to harassment and bullying, and platform integrity harms stemming from misleading AIGC) if our community did not have clear policy guidance for our creators and users on acceptable synthetic media use. We wanted to create clear norms around synthetic media on-platform by developing a policy that would empower creators to explore the creative potential of AI transparently, while protecting against the risk of misleading viewers if they weren't aware that content was edited or created with AI. As part of our policy development process, we engaged with experts including members of our Safety Advisory Committee, WITNESS, and MIT's Dr. David Rand. Dr. Rand studies how viewers perceive different types of AI labels. Dr. Rand's research helped guide the design development of our AI-generated content labels.

In April 2023, we launched a policy that required creators to disclose realistic synthetic media. Our policy asked users to disclose their use of synthetic media from its launch in a method of their choosing (e.g. a sticker, a caption). However, to make it even easier and to support consistent context clues for viewers, we subsequently worked with our Trust & Safety Product team to develop a toggle that users can activate to apply a label to their own AIGC. A significant challenge here was deciding where to draw the line around which uses of AI tools in user-generated content (UGC) we would ask users to proactively label. This was an important set of decisions to make

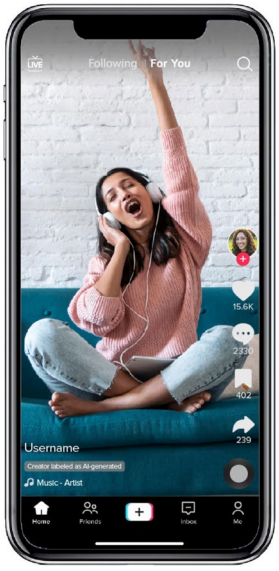

IMAGE 1.
A screenshot depicting the label that creators are able to add to their AIGC.

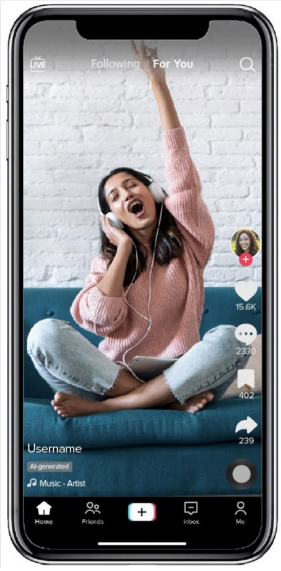

IMAGE 2.
A screenshot depicting the label applied to in-house effects

because we didn't want to contribute to viewer fatigue or undermine the impact of the label by applying it to too wide a range of content — for example, to any minor edits made with AI tools, or filters that don't make the underlying image unrecognizable. Ultimately, we landed on requiring users to label all realistic uses of AI, and encourage them to label anything that has been wholly generated or significantly edited. This excludes minor edits. While we encourage users to label all AIGC, we will remove unlabeled, realistic-looking AIGC in line with our policies.

3 Objective

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

TIKTOK'S RESPONSE

Our synthetic media policy and disclosure tools seek to empower creativity while providing context for viewers that reduces the risks of being harmfully misled by AIGC, and improves trust and transparency about how content is made.

Specifically, our policy has the goal of using disclosure tools to increase transparency and reduce harm to users on the platform, while reducing the potential for real world harm through misinformation, deceptive behaviors like impersonation, or harassment and bullying. This was the goal at the outset of our policy development process and it did not change over the course of the process.

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.

TIKTOK'S RESPONSE

Several of the principles in the PAI Framework relate directly to the values we instill in policy development. As both a Builder of synthetic media creation tools and a Distributor of synthetic media, in the sense that we empower users to share content with each other, we aim to prevent harmful uses of synthetic media and believe strongly in promoting transparency and disclosure.

The first step in our policy development was to identify what uses of synthetic media could be harmful and determine how to work those into our policy in a way that was specific enough to be useful while broad enough to be able to encompass new uses of synthetic media as they arise. While harm reduction is not explicitly a PAI Framework principle, Appendix B of the PAI Framework identifies numerous harms associated with synthetic media which we incorporated into our policy development. Those that seemed most immediately relevant to platforms like TikTok related to information ecosystem harms (misinformation, particularly related to politics and elections) and protecting our community (bullying, harassment, and non-consensual depiction). However, we continue to do risk assessments by monitoring on- and off-platform trends related to AIGC and will update our policy accordingly as the technology evolves.

Disclosure is a major component in our Community Guidelines around synthetic media — we ask users to disclose their use of synthetic media in their content just as we build disclosure into our own AI Effects. For example, TikTok Effects that completely generate or significantly edit content with AI are either labeled as "AI-generated", or have AI disclosed in the Effect name so that it's visible to the viewer through the effect anchor (e.g., AI Portrait). The toggle users can turn on when they post synthetic media adds a label to their content that says, "Creator labeled as AI-generated". Users can tap this label to learn more about AI-generated content. We will continue to experiment with the best way for both the platform and users to disclose as the relevant technology evolves.

Our disclosure efforts cannot be separated from our efforts to be transparent with our users about what content is created with AI, and to provide users with information and guidance around why we label AIGC, and why we ask them to do the same. To further help them understand how to contextualize AI-generated content, we also released detailed Help Center guidance and an in-app explainer that pops up when users tap the label. We continue to roll out entertaining educational content about labeling AIGC. This context is always clearly visible on the content itself, and videos with TikTok effects automatically show an anchor describing the effect name.

5 Obstacles

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

TIKTOK'S RESPONSE

The biggest obstacle that we faced when developing our policy and disclosure mechanisms was on defining precisely what fell in and outside the scope of our policy, and what type of content would require disclosure. Because there is no clear industry standard on how to define rapidly evolving synthetic media, and because the technology is relevant to different platforms and used in different and unique ways, it's natural that this challenge could not be solved by the PAI Framework's broader definition alone. Nevertheless, it might be useful as a community of practice to develop some basic considerations around what type of manipulation by AI might meet the threshold for requiring disclosure, given that this is a key norm that the PAI Framework wants to socialize. This would also help all three types of groups identified in the PAI Framework (**Builders**, **Creators**, and **Distributors**) implement the PAI Framework's recommendations.

6 Benefits

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

TIKTOK'S RESPONSE

Engagement during the development of PAI's Framework was helpful with regard to our harm-focused approach to policymaking, as it provided examples of harm related to malicious synthetic media use that helped guide our thinking around acceptable and non-acceptable uses that we worked into our Community Guidelines. Additionally, keeping our commitments to transparency, harm reduction, and disclosure in mind helped us guide the development of new AI Effects, both with regards to harm reduction and disclosure, as well as ongoing updates to our policy language.

7 Conclusion/Key Takeaways

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

TIKTOK'S RESPONSE

As discussed above, the biggest challenge in our policy development process was defining what fell in and out of scope. This is particularly important for platforms like TikTok that empower people to edit and share content as a way of expressing their creativity. We understand that while synthetic media may cause harm, it's also a tool that provides new avenues for creative expression for our users and positive uses we welcome. We also want to make our disclosure requirements easy for users to understand, and to be able to consistently apply to our own in-house effects and to UGC, which requires a consistent definition of what media is "synthetic enough" that it warrants disclosure.

From the perspective of a platform, if **Builders** (as defined in the PAI Framework) would implement more content provenance/metadata or watermarking techniques in their models, it would greatly benefit our detection and labeling efforts. We're committed to working with the industry through content provenance partnerships and are actively working to implement provenance standards.