PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

How AI video startup Synthesia is scaling up content moderation at point of creation

synthesia

This is Synthesia'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.

SYNTHESIA'S RESPONSE

Synthesia is a global AI video creation platform for enterprises — making video production simple and intuitive without the need for cameras or studios.

More than 50,000 businesses, including more than half of the Fortune 100, use Synthesia's technology to create training, instructional, and product marketing videos in a matter of minutes. Synthesia clients have generated over 12 million videos in the past few years, replacing text-based communications with engaging video.

Synthesia users simply type in text, select from over 200 diverse AI avatars, and choose from more than 130 languages to produce videos that speak to a variety of company audiences. The avatars are generated based on video and voice inputs produced by Synthesia's AI models, presenting the scripts provided by the platform's users.

Powered by generative AI, Synthesia is transforming physical video production into an entirely digital process that will enable creators to bring their ideas to life, from training videos to Hollywood-quality productions.

Synthesia's team of over 300 is headquartered in London, with offices in Copenhagen, New York, Amsterdam, Munich, and employees distributed around the globe.

2 Challenge

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

SYNTHESIA'S RESPONSE

Because of how powerful generative AI can be and how quickly the technology is evolving, Synthesia has been focused on trust and safety since day one. What's more, our client base, including some of the largest enterprises in the world, requires us to ensure the product is safe and that we do everything possible to eliminate misuse.

The company was founded on an ethical framework of consent, control, and collaboration — these principles (internally known as the 3Cs) guide both product and business decisions.

The implementation of the consent and control principles relates to both policy design and to technical execution. We explore both in this case study, focusing on consent when it comes to creating AI avatars and control as it pertains to content moderation and how people interact with our platform. Although the third principle of collaboration is not in scope for this case study, it is equally important as it allows us to achieve the first two. Our collaborative efforts include engaging with regulatory and industry bodies to champion the formulation of robust AI policies and regulations. In addition to being launch supporters of Partnership on AI's (PAI) Synthetic Media Framework, we work alongside companies such as Adobe, Nvidia, and Microsoft, as part of the Content Authenticity Initiative to develop an open industry standard for content authenticity and provenance.

As a rapidly growing company, operationalizing consent and control comes with practical challenges. For example, we chose the principle of consent because we believe people have the right to know and decide how their likeness will be used. As a result, we made an early decision to only create an AI avatar with the explicit consent of the person, which is different from the approach of other generative AI startups that have allowed for synthetic media to be created of public figures, such as actors or musicians, without these figures' consent.

Synthesia offers two types of avatars: stock avatars which are generated based on a composite of real human actors, and custom avatars which are created based on an individual's likeness (a digital twin). In order for anyone's likeness to become a Synthesia avatar, we need footage of them giving verbal consent before we can begin that process. We also give people the ability to opt out of either contributing to a stock avatar or of using their own, and we guarantee their data and likeness will be entirely deleted from our databases.

To implement the principle of consent for custom avatars, we also built a verification process similar to

"know your customer" (KYC) in the banking industry that ensures the owner of the custom avatar can decide who uses their avatar, when, and how. As an early-stage startup, allocating engineering effort for verification was a hard choice, but one that the co-founders deemed an essential investment into the long-term success of the company.

Ultimately, the most important decisions we've made relate to the early investment and continuous evolution of our content moderation systems. Early-stage startups are resource-constrained, so deciding to invest time and money in content moderation early on meant there were other things we couldn't invest in initially. It ended up being the right decision for Synthesia, as the content guardrails we have in place ensure our product is safe and effective for our core use case of enabling video creation for enterprises.

Our strategy with content moderation has been to implement detection capabilities at the point of creation, which was a new approach at the time. Until recently, most content moderation has happened at the point of distribution: a user of digital creation tools could create content without any restrictions, but would then be moderated when distributing the content using an internet service such as a social media platform. The internet service would analyze the content once it was live and then make a decision whether it should be removed or not — but by that point, the content could've received widespread distribution, making this process somewhat ineffective and reactive.

By moderating content at the point of creation, we can better ensure that our policies are adhered to and we can catch more harmful content before it ever generates, making our approach more proactive.

3 Objective

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

SYNTHESIA'S RESPONSE

There is a natural link between consent and content moderation, yet it is rarely explored. Creating AI avatars only with explicit consent makes us more effective at mitigating the harms identified in Appendix B of PAI's Framework. In addition, our objectives with implementing guardrails around consent and content moderation were to adopt a sustainable and ethical business model that fostered a culture of responsible research and product development. As a result, our goals were to:

1. **Build the most trusted AI video platform for the enterprise:** Our business model is aligned with our ethical stance. As our core business is to serve everyone from Fortune 100 clients to independent creators and small businesses, we have no incentive to enable bad faith actors and believe that access and control need to be balanced. Therefore, as **Builders** of business-to-business (B2B) technology and infrastructure, our business model is different from other companies building business-to-consumer (B2C) AI video solutions. Synthesia also introduces an additional application layer which abstracts away the foundational AI model from the user. Rather than interfacing directly with our models, users create videos by simply providing a script and any additional visuals — and our platform automates the rest of the content creation process in a more deterministic way, reducing the potential risks associated with generative AI such as hallucinations, regurgitation, accidental leaks, or malicious attacks.
2. **Make video accessible:** Our mission at Synthesia is to make video easy. The vast majority of our users greatly benefit from the lower barriers of entry that AI video enables. Without the need for cameras, studios, special software, or special editing skills, Synthesia empowers users to generate video just by typing text. By mitigating misuse through consent and control, we ensure this large majority can benefit from AI video long-term. Ensuring that they have a safe platform is a core priority.
3. **Contributing to a healthy ecosystem:** While no single company can affect systemic change, we believe that collaborating with the wider ecosystem across the full content creation and distribution value chain leads to a healthy media environment. As we continuously improve our moderation systems, collaborating with other responsible platforms is a key source of insight. Similarly, Synthesia also contributes lessons learned from implementation and iteration of policies. Being part of building this healthy ecosystem was Synthesia's

main reason for joining PAI's Framework as a founding supporter. That is also why we apply content moderation at the point of creation to videos created with Synthesia. We believe this approach will keep the platform safe for all of us to create professional videos and contribute to a healthier information ecosystem — by reducing the spread and reach of harmful synthetic media. While we are setting the highest possible standards for the use of our technology, we also think we have a responsibility to educate people on how to use generative AI responsibly, respectfully, and creatively. Our content moderation policies define two types of content that help us identify what is outside the scope of a healthy information environment:

a. *Prohibited content* refers to topics that are completely prohibited from use on the Synthesia platform. Prohibited content includes content that could depict acts of violence towards specific individuals or groups, hate speech, or content involving illegal goods or services. This is to ensure that our AI avatars remain free from associations with specific topics that may not align with their intended use.

b. *Restricted content* refers to topics that are restricted to custom avatars only. Restricted content also includes content that could depict misinformation or information of a deceptive nature. This is not only to protect the individuals who represent our stock avatars, but also to create well-defined boundaries between educating an audience on a certain topic and misleading them. For example, we may allow informational, educational, or promotional content about working abroad but we prohibit misleading, unverified, or unauthorized information or advice about visa requirements or the promotion or sale of forged, counterfeit, or fraudulent travel documentation.

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.

SYNTHESIA'S RESPONSE

Our 3Cs of consent, control, and collaboration clearly map to PAI's Framework principles and to the harms it aims to mitigate. In fact, we joined the Framework as a consequence of our ethics principle of collaboration. Our reason for this principle is that no single stakeholder can enact system-level change without public-private collaboration. Ecosystem-wide initiatives led by groups such as PAI are essential to success.

Synthesia started implementing responsible AI practices before the final Framework text was agreed upon. In certain instances, we made the decision to go beyond its guidance, implementing narrower interpretations or developing stricter policies around consent and control that more readily relate to Synthesia's context. These decisions were guided mostly by our core clientele of large enterprises, which tend to mandate stricter conduct requirements than if we had focused on building a pure B2C platform.

For example, PAI's Framework recommends informed consent, which can take many forms. Our approach of explicit consent leaves less space for uncertainty, with clear requirements for verification and no room for compromise for creating someone's AI avatar without their knowledge. We obtain a person's explicit consent in our studio: we ask people participating in the process of creating an avatar to agree for their likeness to be captured and used by Synthesia. We explain how the technology works and how their avatar can be used once it's part of our platform. We also give people the ability to opt out. We respect that decision, ensuring a smooth transition as their data and likeness is removed from our databases.

Of course, our approach may not be the right one for others whose products have different audiences, goals, and purposes. Each company will have to decide what they deem as the right approach to implementing its ethical stance. The nuances of any company's ethical stance will manifest in the actual business and product decisions it makes. The complexity around bridging ethics and execution is exactly why we decided to highlight the challenges Synthesia had to navigate when implementing content moderation systems early, which we will elaborate on next.

5 Obstacles

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

SYNTHESIA'S RESPONSE

Content moderation is an ever-evolving process which involves constantly adapting to new threats by developing new policies or implementing more advanced moderation systems. As a company grows, doing content moderation at scale becomes one of the biggest obstacles to overcome. This is why we decided to invest in content moderation capabilities: at the end of 2023, almost 10% of the company was dedicated to working on trust and safety. Additionally, we operationalized content moderation according to three broad categories:

1. **Obviously harmful content:** Content that is generally prohibited across most platforms. It is by far the least common of the three categories, but also the most straightforward to take action on as it clearly causes harm. Identity-based hate, child endangerment, terrorism, incitement of violence, or graphic content would fall into this category.
2. **Obviously harmless content:** Most enterprise clients use Synthesia to create onboarding and training videos about company processes and products, with no reference to any ambiguous topic. These are obviously harmless and easy to screen with automated systems, if needed. An example would be an instructional video teaching salespeople about their company's new product or process.
3. **The gray zone:** Content that, depending on context, can be either benign or harmful. An example is content related to sexual health or emerging technologies such as cryptocurrencies: while we want to empower clients who create educational videos, the same vocabulary can be misused to misinform or deceive an audience.

Governing the gray zone poses the biggest challenge, both to policy design and technical implementation. The gray zone is also the reason why we believe that a useful moderation setup requires a combination of human moderators, augmented by machine moderation, to handle the first two obvious categories at scale, while also triaging for ambiguity in the third.

POLICY CHALLENGES

Our policies were developed and evolved considering the UK government's safety by design principles and preventative approach. In line with the guidance, we used the following criteria to assess whether a specific piece of content should be allowed to be created with Synthesia:

1. Does the content pose a risk of material, physical, or emotional harm to individuals or a community?
2. Was the content likely created with intention to harm?
3. Was the content likely created to enable criminal acts?
4. Was the content created to educate or inform?

While at first glance these might seem to be straightforward decisions, nuanced implementation of our policies was challenging, as it required identifying every potential edge case to then answer the above questions and to act. Early on, we realized that for our moderation guidelines to be useful to our moderation team, they had to include very specific guidance not only regarding the topic of the content, but also the context in which the topic is discussed.

For example, Synthesia's users may attempt to create sexual health videos to educate and inform their viewers. This is clearly a behavior we aim to reinforce as research has shown that video-based training can be more effective than text. So, we allow content that educates about sexual health, including reproductive health or disease prevention. However, we cannot allow all content related to sexual health. Videos that contain opinions that contradict medical expert consensus would fall under prohibited use of our platform.

Synthesia treats content related to sexual harassment similarly. We allow content that educates people about what sexual harassment is and what steps they can take to identify and report it. However, we would not allow our platform to be used for the purpose of harassing others or misleading them about what harassment is — again, context is important. In one example, our automated systems identified a script from an account that was making references to sexual harassment which triggered a review of the script from our human moderators. Upon closer inspection, we noticed that the script aimed to educate the employees of a small company on how to report harassment to human resources. As a result, we allowed them to create the video.

The same principles can be applied to emerging technologies such as blockchain, cryptocurrencies, or NFTs. If the content produced aims to educate people about the technical aspects of a particular blockchain

technology, then that content is allowed. However, if someone attempts to create a video to deceive people into investing in a cryptocurrency (a practice known as "pump and dump"), it would be against our policies.

In an example from our platform, our moderation system flagged a script from an account that was making references to cryptocurrency trading. When we reviewed the contents of the script, it became clear that it was making claims of guaranteed profits by engaging in cryptocurrency trading, with misleading information about their returns — something that is banned by our content moderation policies related to financial solicitation. We did not allow the user to create the video.

By consequence of the number of content topics on which a user can create videos, and the sheer number of specific permutations in which those topics can manifest, implementing context-specific moderation rules was an iterative process where more nuance was added to cover more and more outliers. Over time, and with the growth of our trust and safety function, we have further reviewed and developed our existing policies. By identifying gaps and working to foster alignment with industry standards, we have improved our internal moderation systems and implemented more nuanced and thorough content moderation policies. Additionally, we expanded our transparency efforts and expanded our content moderation guidelines to improve customer experience and foster trust within our platform. We continuously review and update our systems with new topics and keywords that will trigger human moderators to evaluate a script.

Video as a format adds further complexity to content moderation challenges. Different measures are necessary for videos created with a custom AI avatar, featuring one's own likeness. For example, political opinions can be shared with one's custom avatar, as long as they don't target groups or individuals and are not meant to incite violence or intend to harm. The same political opinions are not permitted, however, in videos featuring stock AI avatars, to mitigate the potential misuse stemming from the anonymity and degree of separation that stock avatars might provide.

6 Benefits

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

SYNTHESIA'S RESPONSE

Since 2017, we have identified over 60 content categories, most of which include gray zone areas. Our Trust & Safety Operations team constantly reviews novel cases and refines existing guidelines. The iterative nature of moderation also means we need to acknowledge that systems never reach perfect efficacy. While that is a reality we need to face, there are many benefits of continual upkeep and iteration.

First, there is no real alternative to continuous improvement. Content moderation is not a static set-and-forget area. Continuous iteration is what makes the system useful. This has broad ramifications, especially for enterprises, where new AI compliance processes and teams might need to be established.

Second, it facilitates a scalable model for human-machine moderation systems. The content categories and specific cases provide an easy decision map for human moderators, improving both moderation accuracy and time.

Lastly, an iterative approach doesn't mean companies need to start from zero. One of the clear benefits of our participation with PAI is the community of experts who co-created PAI's Framework. Both the final text, which includes key areas of harm to consider, as well as the know-how and experience from the co-creation process, contributed to successfully bridging the gap between ethics principles and execution.

7 Conclusion/Key Takeaways

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

SYNTHESIA'S RESPONSE

Applying a KYC-style consent process and content moderation at the point of creation are novel concepts, but ones that should clearly be considered by more generative AI platforms. Like many others, we also believe that the sustained health of the media ecosystem is dependent on such responsible practices.

While we believe all companies involved in the generative AI ecosystem should embrace content moderation, the exact scope of implementation will differ. Content moderation decisions are grounded in the worldview of a certain organization and based on a variety of factors, such as company culture, risk appetite, and customer segment.

We also acknowledge that committing 10% of employees to Trust and Safety can be challenging for platforms that serve other markets. As an enterprise SaaS business, we believe that such high-resource commitment contributes to building a business that is successful in the long-term.

We've decided to implement policies that some might find too stringent, but we believe these terms serve our clients and community best. For example, our KYC-type program allows us to implement additional security measures that help our platform to operate ethically, securely, and in line with its intended purpose. Our onboarding flow is designed to limit inauthentic behavior such as impersonation, identified in Appendix B of the Framework, which is a significant vector for the distribution of dis- and misinformation on the internet. Our content moderation policies also prohibit stock avatars from being used to represent employees or real-life people without disclosure or to impersonate a private or public individual (alive or deceased).

In the coming years, as generative AI matures and proliferates, every company will have to make its own assertions of how to implement responsible AI practices. PAI's Framework and community will be there to support them.