BR Business Roundtable

Safely Harnessing Synthetic Content's Potential

September 2024

Overview

As artificial intelligence (AI) technology becomes more advanced and accessible, AI-generated synthetic content — both beneficial and harmful — will continue to emerge. AI systems are reshaping an already dynamic and complex risk landscape, which necessitates thoughtful guardrails.

Regulatory guardrails should be targeted to effectively address specific risks associated with synthetic content, while encouraging credibility and enabling beneficial use cases. As policymakers consider their approach to address concerns stemming from synthetic content, they should:

- **Consider how synthetic content is used and the corresponding benefits and risks;**
- **Prioritize techniques and standards to identify and authenticate credible content rather than solely relying on labeling synthetic content;**
- **Layer technical and people-centric approaches for maximum effect in safeguards meant to manage risks of synthetic content; and**
- **Ensure organizations have the opportunity to remediate harmful synthetic content.**

Background

Synthetic content is information — such as an image, video, audio clip or text — that has been significantly modified or generated by AI. Generative AI's evolution has made synthetic content an increasingly important issue. As AI is deployed across business functions and sectors, synthetic content will increasingly be integrated into business-to-business and consumer-facing contexts to serve a wide variety of purposes, many of which are or will be widely accepted. For example, AI-generated text can provide first-line customer service for many organizations. AI-generated data, known as synthetic data, also plays a critical role in enabling comprehensive testing without compromising privacy or security. However, AI-generated synthetic content — just like altered or intentionally misleading images, media, etc. created by any other means — can also fuel misinformation, undermine trust in democracy and institutions, be used to perpetrate fraud and cyber attacks, disrupt markets, and cause harm to targeted individuals and groups.

Trust is essential to a safe and healthy society, political environment and economy. Businesses need to be sure that their customers and suppliers are who they claim to be, and individuals need to be confident they can trust online information. A thoughtful regulatory framework is critical for mitigating the potential harm of synthetic content and increasing trust in AI.

Policy Considerations and Recommendations

To design effective policies and guardrails for synthetic content, policymakers should:

Account for context of use

Business Roundtable Recommendation:
Policymakers should adopt risk-based guardrails for synthetic content that are adaptable and protect beneficial uses.

How content is used should be the key consideration for policymakers — not whether content is created or modified using AI. The same AI tools can be used to modify or generate many different types of content for a wide range of purposes. For example, a photo editing algorithm can enhance a blurry image, which does not fundamentally change the meaning of the image, or it can be used to alter images to change their context (e.g., location, individuals present), which may cause misinformation and harm under some circumstances. Similarly, chatbots can create compelling educational content or help direct customers to assistance, or they may be used to create content that is intended to manipulate. Synthetic data and content are also widely used in the business community for internal operations, which are not public-facing and thus do not present the same opportunities for misuse. While the AI technology used to create content in each of these contexts is similar, the content's ability to mislead and potential to cause harm is not.

Policymakers should approach risk mitigation in well-scoped, targeted ways that address real-world use cases and risks. Potential harm is dependent on the context in which synthetic content is created and shared rather than the content itself. Policymakers should focus guardrails on potential harmful outcomes rather than regulating the tools used to create content.

Potential harm is dependent on the **context** in which synthetic content is created and shared rather than the content itself.

When developing guardrails, policymakers should adopt a holistic approach that accounts for the specific risks of a particular use case, as well as the potential benefit. Many factors may contribute to this analysis, including: the content medium (e.g., image, audio, text); whether the content is entirely new or modified from existing content; the extent to which any modifications materially impact the meaning of content; and how public facing the content is likely to be and what audiences are likely to consume it.

Prioritize authenticity

Business Roundtable Recommendation:
Policymakers should support initiatives that seek to validate authentic and credible content, ensuring individuals have sufficient information to identify the source and evaluate the trustworthiness of the content they encounter.

Most concerns around synthetic content stem from a viewer's inability to determine whether the content is trustworthy, rather than how it was created. Detection of synthetic content does not necessarily identify whether that content poses a risk or should be considered credible. For example, public awareness campaigns can leverage synthetic content that is accurate and compelling. The important aspect is that viewers know it comes from a credible source.

Policymaking approaches to synthetic content should focus on demonstrating authenticity and the source of the content, or provenance, through mechanisms that embed these attributes. Provenance indicators are distinct from whether content is synthetic and can help people make their own determinations about its trustworthiness. This is similar to industry approaches around digital identity that have served to combat fraud, improve consumer privacy and advance online security as consumers interact and transact online.

Policymaking approaches also need to consider likely behaviors of bad actors while building public trust. Policymakers should prioritize measures aimed at stopping bad actors rather than simply adding requirements that increase the compliance burden for good actors but are easily avoided by bad ones. Exclusively relying on models or content creators to label synthetic content as a bulwark against misinformation will fail because bad actors could choose to use different models or simply not label their synthetic content. Meanwhile, initiatives focused on provenance and authenticity can inspire public trust by clarifying content sources and their credibility, even after the content is disseminated. The ability to determine content provenance and authenticity is likely to be more useful than knowing whether content is synthetic, though different contexts will be served by different approaches — often in combination.

Policymakers should prioritize measures aimed at stopping bad actors rather than simply adding requirements that **increase the compliance burden** for good actors.

Embrace a combination of technical and people-centric approaches

Business Roundtable Recommendation:
Regulatory guardrails should integrate multiple technical and people-centric approaches to effectively manage the risk of synthetic content to people and society.

Technical approaches to the harms of synthetic content, such as labeling, detection or built-in model safeguards, each have strengths and weaknesses. For example, using labeling to identify synthetic content as it is distributed requires that these markers remain embedded in the content, which is difficult to enforce broadly but can be effective on a centralized platform that is able to control the labeling. Detection of synthetic content can be difficult because models modify or generate content in different ways.

Technical approaches will be **most effective** when paired with people-centric approaches.

These technical approaches will be most effective when paired with people-centric approaches. For example, detection efforts may reveal certain indicators that human fact-checkers and news organizations can use to help determine authenticity. No single method will be workable or effective across all use cases. Instead, complementary approaches can be applied individually and in combination as appropriate.

Give organizations the chance to remove harmful synthetic content

Business Roundtable Recommendation:
If policymakers create frameworks that impose penalties for harmful AI-generated synthetic content, they should ensure that responsible parties are given an opportunity to remediate.

As AI technology proliferates, there will be more malicious and objectionable synthetic content distributed online. Given the perceived risks associated with synthetic content, policymakers may seek to designate certain types as harmful or illegal. This may also inspire the enactment of new liability regimes. Such an approach should be considered carefully since it could severely harm innovation and undermine beneficial use cases.

Liability regimes for model developers may disincentivize the development of new, experimental models, start-ups would struggle to finance legal battles and other businesses may be discouraged by substantial legal risk. This approach may also discourage larger businesses from disseminating their tools more broadly or using synthetic content for established beneficial uses.

If pursued, policymakers should scope liability regimes narrowly, focusing on the most problematic types of synthetic content. Criminal penalties are appropriate for bad actors who maliciously post synthetic content, such as perpetuating cyber fraud and other crimes. Meanwhile, actors who make good-faith efforts to remove harmful content should not face immediate penalties.

Actors who make **good-faith efforts** to remove harmful content should not face immediate penalties.

In the United States, internet platforms and intermediaries already remove harmful and illegal content, often after they are notified that it has been shared (i.e., a notice-and-takedown process). This notice-and-takedown approach has been successful because it allows responsible intermediaries to address concerns through a predictable and well-defined process, while providing remedies against actors who fail to respond. To the extent that any new liability regimes are imposed, an organization should only be held liable and face legal repercussions if it fails to remove the content once appropriately notified.

Conclusion

Synthetic content has a wide variety of beneficial purposes but also has the potential to harm individuals, organizations and ecosystems if it is used to manipulate and misrepresent. As these technologies become widely accessible, policymakers should develop risk-based, context-driven and flexible guardrails that appropriately safeguard against these risks while protecting beneficial uses. The most promising technical approach may be creating methods to verify the authenticity and provenance of content, paired with people-centric approaches. These guardrails should reflect the importance of innovation in AI while emphasizing the significance of ensuring information credibility and authenticity.