Blogs
October 17, 2023
Generative AI: How Existing Regulation May Apply to AI-Generated Harmful Content

Among the many open questions about large-language models (LLMs) and generative artificial intelligence (AI) are the legal risks that may result from AI-generated content.

While AI-specific regulation remains pending and continues to develop in jurisdictions around the world, the following article provides a high-level summary of illegal and harmful content risks under existing law, as well as mitigations that companies may wish to consider when developing baseline models and consumer-facing generative AI tools. (For copyright and intellectual property (IP)-related issues, see Perkins Coie Updates.)

Content Risks

Many content and communication platforms—such as social media platforms—engage in processes to find, block, and remove user content that is potentially harmful to others. This content moderation and related efforts to promote online safety are challenging and complex, particularly at scale. In recent years, lawmakers in the United States and around the world have increasingly sought to regulate how providers reduce the availability of illegal and harmful content. Companies developing or launching generative AI tools may wish to consider how the following categories of existing and pending regulation may apply to their services.

Consumer protection laws. The U.S. Federal Trade Commission (FTC) has made it clear that the FTC Act's prohibition on deceptive or unfair conduct applies to generative AI tools. Such tools can be exploited for phishing scams, identify theft, deepfakes, or the creation of other content that misleads, defrauds, or harms individuals. The FTC has warned that chatbots can be used to generate phishing emails, fake websites, fake posts, fake profiles, fake consumer reviews, and malware, as well as to facilitate imposter scams, extortion, and financial fraud. In a blog post on AI, the FTC encouraged companies to take reasonable precautions before a product hits the market, noting that it has taken action against companies that distributed potentially harmful technologies without making reasonable efforts to prevent consumer injury.

Child safety and age-appropriate content. Without sufficient guardrails, generative AI tools run the risk of accelerating cyberbullying, harassment, sextortion, and the proliferation of other illegal and harmful content, which could lead to investigations and regulatory enforcement. Providers of AI systems will likely need to take steps to comply with child pornography and child sexual exploitation laws, including provisions on computer-generated images. Providers should also implement policies to prevent their tools from being used to facilitate behaviors such as grooming that advance exploitation. Such providers may also need to navigate the obligations imposed by many jurisdictions to report this content.

The FTC and several U.S. states are also carefully scrutinizing how online services may expose kids and teens to other harmful content. Indeed, many states have already enacted laws that address kids' safety online. Other jurisdictions around the world, including Australia, the EU, Ireland, India, and the UK, have also implemented requirements related to age-inappropriate content (such as violent or drug-related content) that may extend to generative AI tools.

Geopolitical and cross-cultural risks. Providers may wish to consider how generative AI output could implicate geopolitical sensitivities and associated legal risks. Several markets have laws that seek to address content that is perceived to undermine or threaten national security, including alleged foreign interference or content that promotes locally designated terrorist organizations. In addition, as generative AI tools expand to include more languages, providers should consider the complexities and risks associated with usage variation across dialects. For example, an innocuous term in one dialect can be hate speech in another. Providers should consider any risks that may arise in connection with datasets that rely on a dominant language or dialect to ensure that they are effectively guarding against misuse or abuse.

Other content regulation laws. Generative AI output may also contravene other existing laws on harmful online content, as defined by jurisdictions in the United States and abroad. In addition to the child pornography and terrorism laws noted above, certain U.S. and foreign jurisdictions regulate obscene content, deepfakes, nonconsensual nudity, self-harm and suicide, sexual services, violent material, and misinformation, among other content categories. Providers will need to guard against the risk of generative AI output being deliberately used to create harmful or illegal content. Content moderation laws that mandate takedown, transparency reporting, and risk assessment practices, such as the EU's Digital Services Act, may also apply to content generated by AI tools that are embedded into online services.

More recent regulation is addressing generative AI output head on. The Australian eSafety Commissioner recently registered a binding code applicable to search engine services that will include content and safety requirements for generative AI tools. A recently adopted Texas law on kids' privacy and safety excludes search engines that create "harmful content" from its general exemption for search engines and cloud services. And the most recent draft of the European Union's AI Act would require testing and analysis (including by independent experts) to identify and mitigate reasonably foreseeable risks to health, safety, fundamental rights, the environment, democracy, and the rule of law. The AI Act would require providers of these models to design, develop, and train the model in such a way as to ensure that they do not generate content that violates EU law. We anticipate seeing more generative-AI-specific content regulation in the months ahead.

Defamation claims. Generative AI models' use of statistical inferences can at times produce incorrect information known as hallucinations. While the elements of defamation claims vary around the world, providers of generative AI tools should be aware that hallucinations may lead to defamation claims, particularly if the model's false statements arguably result in harm to an individual's reputation.

Potential Mitigations

Providers may wish to consider various controls and governance measures to mitigate the above risks. Those developing foundation models have an opportunity to mitigate bias at the outset through careful curation of datasets. Notably, the FTC has indicated that purveyors of generative AI tools should (1) consider the risk of misuse or harm at the design stage, (2) take all reasonable precautions before launching a product, and (3) avoid overreliance on post-release detection.

Those deploying generative AI content tools may wish to consider the following mitigations:

- **Red teaming and impact assessments.** Red teaming exercises and impact assessments allow providers to anticipate and address potential content and safety risks before they arise. The National Institute of Standards and Technology's (NIST) AI Risk Management Framework emphasizes the importance of having policies and practices in place to solicit and integrate feedback from a diverse group of internal and external stakeholders. This proactive approach enables the development of more trustworthy systems and can mitigate legal risks.
- **Standardized and transparent content moderation.** Providers should also consider developing defaults and hard bounds through content policies. To mitigate any consumer protection or content regulation risks, such policies should be enforced according to transparent, proportionate, and standardized review processes.
- **Audits.** Robust risk mitigation should also include post-release controls. Audits allow for internal and external evaluation of content risks and facilitate transparency and consumer confidence about efforts to mitigate such risks. The White House's Blueprint for an AI Bill of Rights recommends that AI systems be designed to allow for independent evaluation of the system's safety and effectiveness, including by third-party auditors.
- **Defamation insurance.** Given the risk that generative AI hallucinations may lead to defamation claims, generative AI providers may wish to assess whether these claims could be covered under their liability policies, including their commercial liability, management liability, cyber liability, or directors and officers liability policies. Where applicable, a generative AI provider may also be covered as an "additional insured" through a contract arrangement with another entity that has liability coverage.

Takeaway

The explosive growth of LLMs and generative AI tools is coinciding with a wave of regulatory activity related to online safety and content moderation across the United States and abroad. While the extent of liability exposure remains uncertain, providers may wish to consider potential risks early in the product development cycle. An important threshold question to many claims will be the extent to which generative AI content will be protected by Section 230 of the Communications Decency Act and other safe harbors for third-party content around the world. Recent federal legislation in the United States proposes a limitation on immunity under the Communications Decency Act in connection with generative AI, and the issue will likely continue to be hotly contested, at least until AI-specific regulation is enacted.

Authors

Explore more in

Privacy & Security