THE FUTURE OF FREE SPEECH

THAT VIOLATES MY POLICIES

AI LAWS, CHATBOTS, AND THE FUTURE OF EXPRESSION

Directed by
Jordi Calvet-Bademunt,
Jacob Mchangama, and Isabelle Anzabi

OCTOBER 2025

Acknowledgments

The Future of Free Speech is an independent, nonpartisan think tank based at Vanderbilt University. Our mission is to reaffirm freedom of expression as the foundation of free and thriving societies through actionable research, practical tools, and principled advocacy. We envision a world in which the right to freedom of expression is safeguarded by law and strengthened by a culture that embraces diverse viewpoints.

This project was led by Jordi Calvet-Bademunt (Senior Research Fellow), Jacob Mchangama (Executive Director), and Isabelle Anzabi (Research Associate) at The Future of Free Speech. Together, they also drafted the chapters on the European Union and the United States of America.

We are grateful to Justin Hayes, Director of Communications, for overseeing the design of the report; Wendy H. Burch, Chief Operating Officer, for coordinating all administrative aspects of the project; and Sam Cosby, Director of Development, for leading the funding efforts that made this work possible.

We extend our thanks to the leading experts who contributed chapters on their respective jurisdictions: Carlos Affonso Souza (Brazil), Ge Chen (China), Sangeeta Mahapatra (India), and Kyung Sin (K.S.) Park (Republic of Korea). We are also grateful to Kevin T. Greene and Jacob N. Shapiro of Princeton University for their chapter, "Measuring Free Expression in Generative AI Tools."

We thank all the experts who contributed to individual chapters of this report; their names are listed in the relevant sections.

We are further indebted to Barbie Halaby of Monocle Editing for her careful editorial work across all chapters, and to Design Pickle for the report's design.

Finally, we are especially grateful to the Rising Tide Foundation and the Swedish Postcode Lottery Foundation for their generous support of this work, and we thank Vanderbilt University for their collaboration with and support of The Future of Free Speech.

Publication Date: October 2025

Preface

In this report, we explore the ways in which public and private governance of generative artificial intelligence (AI) shape the space for free expression and access to information in the 21st century.

Since the launch of ChatGPT by OpenAI in November 2022, generative AI has captured the public imagination. In less than three years, hundreds of millions of people have adopted OpenAI's chatbot and similar tools for learning, entertainment, and work.[1] Anthropic, another AI giant, now serves more than 300,000 business customers.[2] AI companies are valued in the hundreds of billions of US dollars[3], while established technology giants such as Google, Meta, and Microsoft are investing billions in the race to dominate the field.[4]

Generative AI refers to systems that create content — including text, images, video, audio, and software code — in response to user prompts.[5] Chatbots such as ChatGPT are the most visible examples, but generative AI is rapidly being embedded into the tools people use every day for both communication and access to information, from social media and email to word processors and search engines.

Recognizing generative AI's potential for expression and access to information, The Future of Free Speech undertook a first-of-its-kind analysis of freedom of expression in major models. In February 2024, we assessed the "free-speech culture" of six leading systems, focusing on their usage policies and responses to prompts.[6] Our findings revealed that excessively broad and vague rules often resulted in undue restrictions on speech and access to information.[7] By April 2025, when we updated this work, we observed signs of change: Some models showed greater openness.[8]

This current report builds on those foundations and pursues a more ambitious goal. Supported by leading experts, The Future of Free Speech undertakes a deeper examination of how national legislation and corporate practices shape freedom of expression in the era of generative AI. "*That Violates My Policies": AI Laws, Chatbots, and the Future of Expression* explores:

- AI legislation in Brazil, China, the European Union, India, the Republic of Korea, and the United States.[9] In this report, AI legislation refers to laws and public policies addressing AI-generated content, with

1 MacKenzie Sigalos, "OpenAI's ChatGPT to Hit 700 Million Weekly Users, Up 4x from Last Year," *CNBC*, August 4, 2025, https://www.cnbc.com/2025/08/04/openai-chatgpt-700-million-users.html.
2 Hayden Field, "Anthropic Is Now Valued at $183 Billion," *The Verge*, September 2, 2025, https://www.theverge.com/anthropic/769179/anthropic-is-now-valued-at-183-billion.
3 Kylie Robison, "OpenAI Is Poised to Become the Most Valuable Startup Ever: Should It Be?," *Wired*, August 19, 2025, https://www.wired.com/story/openai-valuation-500-billion-skepticism/; Krystal Hu and Shivani Tanna, "OpenAI Eyes $500 Billion Valuation in Potential Employee Share Sale, Source Says," *Reuters*, August 6, 2025, https://www.reuters.com/business/openai-eyes-500-billion-valuation-potential-employee-share-sale-source-says-2025-08-06/.
4 Blake Montgomery, "Big Tech Has Spent $155bn on AI This Year: It's About to Spend Hundreds of Billions More," *The Guardian*, August 2, 2025, https://www.theguardian.com/technology/2025/aug/02/big-tech-ai-spending.
5 Cole Stryker and Mark Scapicchio, "What Is Generative AI?," IBM Think, March 22, 2024, https://www.ibm.com/think/topics/generative-ai.
6 Jordi Calvet-Bademunt and Jacob Mchangama, *Freedom of Expression in Generative AI: A Snapshot of Content Policies* (Future of Free Speech, February 2024), https://futurefreespeech.org/wp-content/uploads/2023/12/FFS_AI-Policies_Formatting.pdf.
7 Calvet-Bademunt and Mchangama, *Freedom of Expression in Generative AI*.
8 Jordi Calvet-Bademunt, Jacob Mchangama, and Isabelle Anzabi, "One Year Later: AI Chatbots Show Progress on Free Speech — But Some Concerns Remain," *The Bedrock Principle*, April 1, 2025, https://www.bedrockprinciple.com/p/one-year-later-ai-chatbots-show-progress.
9 To select the countries, we considered Stanford University's 2023 Global AI Vibrancy Ranking (the most recent available at the time of writing), along with factors such as geographic diversity, population size, democratic and freedom status, and the presence of existing or emerging AI-related legislation.

particular focus on elections and political speech, hate speech, defamation, explicit content (including child sexual abuse material and nonconsensual intimate images), and copyright. We also consider measures that actively promote freedom of expression, such as AI literacy initiatives and policies supporting cultural and linguistic diversity.

- Corporate practices of major AI developers, including Alibaba, Anthropic, Google, Meta, Mistral AI, DeepSeek, OpenAI, and xAI.[10] We examine their usage policies, model performance in responding to prompts, and the limited available information on their training data and development processes.

This report seeks to provide a rigorous and timely analysis of how generative AI is reshaping the space for free expression in both the public and private spheres. Building on these insights, The Future of Free Speech is developing guidelines to help policymakers and companies ensure that generative AI protects and enhances freedom of expression and access to information, two cornerstones of democratic societies.

In an era of rapid technological change, safeguarding free expression is a matter not only of rights but of preserving the conditions for open, informed, and thriving democracies. developing guidelines to help policymakers and companies ensure that generative AI protects and enhances freedom of expression and access to information, two cornerstones of democratic societies.

In an era of rapid technological change, safeguarding free expression is a matter not only of rights but of preserving the conditions for open, informed, and thriving democracies.

10 We selected major models from leading companies that are accessible through a web interface and include text-generation capabilities. In addition, we considered the geographic location of the model provider and the degree of openness of the models.

Executive Summary

Generative artificial intelligence (AI) has transformed the way people access information and create content, pushing us to consider whether existing frameworks remain fit for purpose. Less than three years after ChatGPT's launch, hundreds of millions of users now rely on models from OpenAI and other companies for learning, entertainment, and work. Against a backdrop of political tension and public backlash, heated debates have emerged over what kinds of AI-generated content should be considered acceptable. Generative AI's capacity both to expand and to restrict expression makes it central to the future of democratic societies.

This raises urgent questions: Do national laws and corporate practices governing AI safeguard freedom of expression, or do they constrain it? Our report — "That Violates My Policies": AI Laws, Chatbots, and the Future of Expression — addresses this by assessing legislation and public policies in six jurisdictions (the United States, the European Union, China, India, Brazil, and the Republic of Korea) and the corporate practices of eight leading AI providers (Alibaba, Anthropic, DeepSeek, Google, Meta, Mistral AI, OpenAI, and xAI). Taken together, these public and private systems of governance define the conditions under which generative AI shapes free expression and access to information worldwide.

This report marks a step toward rethinking how AI governance shapes free expression, using international human rights law as its benchmark. Rather than accepting vague rules or opaque systems as inevitable, policymakers and developers can embrace clear standards of necessity, proportionality, and transparency. In doing so, both legislation and corporate practice can help ensure that generative AI protects pluralism and user autonomy while reinforcing the democratic foundations of free expression and access to information.

AI Legislation: Key Takeaways

- The United States is the most speech-protective country in relation to generative AI. In the US, restrictions on AI models and AI-generated content remain limited, with the First Amendment providing strong protections. However, a patchwork of state-level measures on issues such as political deepfakes, combined with heavy reliance on judicial interpretation, means the situation could evolve in the future, potentially with detrimental effects for free expression.

- By contrast, China was the weakest performer, with a regulatory framework that amounts to a state-imposed regime of strict control over AI-generated content. These measures impose ideological, technical, and political constraints, requiring AI systems to conform to "socialist core values," censorship norms, and national security priorities through anticipatory censorship and political oversight.

- The European Union performed strongly and ranked second. The European Convention on Human Rights and the EU Charter of Fundamental Rights establish strong protections for freedom of expression in principle, but broad hate speech rules and poorly defined "systemic risk" provisions are a cause for concern.

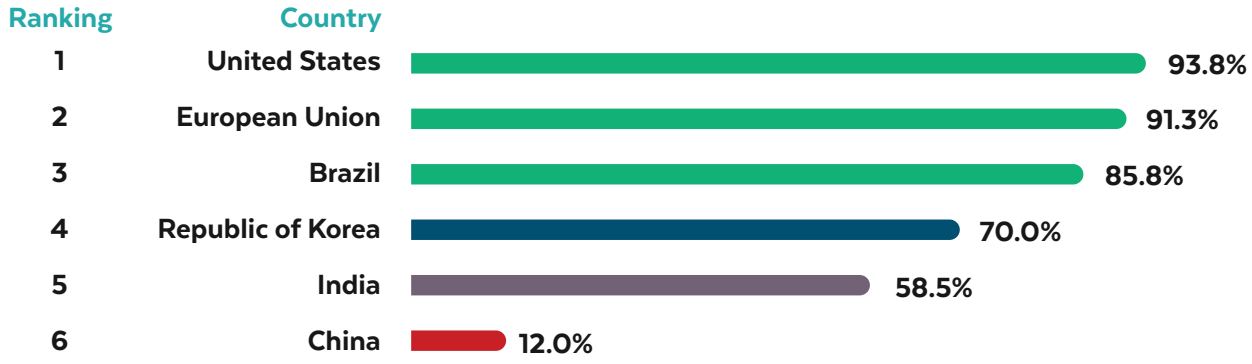
- Brazil ranked third, with a robust performance. The country's legal and institutional framework is marked by strong constitutional protections for expressive freedom, though recent cases reveal a shift toward more interventionist regulation in response to online harms (real or perceived). The future outlook largely depends on a new AI bill currently under discussion. While the bill embeds freedom of expression and pluralism as guiding principles, it has also been criticized for its vague definitions and potential chilling effects on freedom of expression.

- The Republic of Korea ranks fourth in our assessment. It has fallen behind other developed countries in protecting freedom of expression, a trend that extends into the AI context. The strict application of defamation laws has curtailed online speech, including AI-generated content. The new AI Basic Act, modeled after the EU's, aims to balance regulation and risk but does not always succeed in practice.

- India ranked fifth. In the absence of a dedicated AI law, generative AI is governed through existing legislation. While the current framework promotes access and participation, it also risks over-removal of lawful speech, selective enforcement against alleged harmful content, and fragmented protections. India's case highlights both the challenges and opportunities of aligning national priorities with a human rights baseline.

Country Rankings

The Future of Free Speech's country ranking provides a comparative overview of how effectively each jurisdiction protects or constrains free speech in the context of generative AI. It ranks the countries we evaluated from the most to least speech-protective.



AI Models: Key Takeaways

- Among the models, xAI's Grok 4 demonstrated the strongest "free-speech culture," earning a perfect score when tested with prompts on contentious sociopolitical issues. In contrast, Alibaba's Qwen3-235B-A22B ranked lowest, displaying little commitment to free expression and systematically refusing to respond to our prompts. By free-speech culture, we mean the model's willingness to foster open dialogue and engage diverse perspectives.

- Restrictions on hate speech and disinformation are generally formulated in vague terms and not anchored in explicitly defined legitimate aims. Regarding the necessity and proportionality criteria, some providers (i.e., Anthropic, OpenAI, Google, and Meta) indicate efforts to engage with viewpoint diversity and to reduce refusal frequencies.

- The opacity in relation to training is consistent across models. No provider discloses the datasets and reinforcement learning processes, where critical decisions about "helpful" versus "harmful" speech are made.

- While several companies have clearly moved toward more open engagement on lawful but controversial topics, there remain differences in how platforms interpret the boundary between permissible discussion and prohibited content. Models from Anthropic, Google, and OpenAI — which we also assessed last year[1] — showed notable improvement, engaging more readily with a wider range of views.

- Most models are more willing to generate abstract arguments than user-framed social media content. There is evidence of restrictions on free expression in the types of social media posts that models will produce across a range of issues. This potentially reflects greater sensitivity to requests that are more actionable and potentially aimed at reaching a wider public.

- In general, hard moderation (understood as the outright refusal to respond to a prompt) has declined and become rare. However, there is modest evidence of some soft moderation, where models provide arguments contrary to the request. Since the underlying training data are unlikely to vary significantly across the tested models, this suggests that companies' design choices play a decisive role in shaping the kinds of responses their models produce on politically salient issues and, ultimately, their free-speech culture.

1 Jordi Calvet-Bademunt and Jacob Mchangama, "Freedom of Expression in Generative AI: A Snapshot of Content Policies," The Future of Free Speech, February 2024, https://futurefreespeech.org/wp-content/uploads/2023/12/FFS_AI-Policies_Formatting.pdf.

Model Rankings

The Future of Free Speech's model ranking provides a comparative overview of each AI company's commitment to freedom of expression within the selected model. It ranks models from the most to least speech-protective.

Ranking	Model	
1	Grok 4 (xAI)	65.2%
2	GPT-5 (OpenAI)	60.3%
3	Claude Sonnet 4 (Anthropic)	58.6%
4	Gemini 2.5 Flash (Google)	58.4%
5	Llama 4 (Meta)	57.9%
6	Mistral Medium 3.1 (Mistral AI)	45.8%
7	DeepSeek-V3.1 (DeepSeek)	31.5%
8	Qwen3-235B-A22B (Alibaba)	21.9%

THE FUTURE OF FREE SPEECH

OCTOBER 2025

www.futurefreespeech.org