From “Googling” to “Asking ChatGPT”: Governing AI Search

AI FORENSICS

November 2025

R. Buse Çetin Natalia Stanusch Marc Faddoul

CREDITS

The first version of the report was published on December 3, 2025. The information in the report was updated multiple times due to changes in products and systems in the scope of this report, as well as regulatory developments. Given the evolving and pressing importance of this subject matter, we might release further updates to this report.

Authors: Raziye Buse Çetin, Natalia Stanusch, Marc Faddoul

Reviewers: Martin Degeling, Salvatore Romano

We extend our gratitude to Beatriz Botero Arcila and Laureline Lemoine for their invaluable insights and feedback on this report.

Graphic & Brand Design: Denis Constant / Ittai Studio / https://ittai.co/

Email: info@aiforensics.org
Social Media: LinkedIn | X | Bluesky

From “Googling” to “Asking ChatGPT”: Governing AI Search

Table of Contents

How to Read This Report

This report presents arguments related to empirical research, platform moderation, the regulatory landscape, and the relevant technical background of both search engines and AI search applications. Since we bring together platform studies, empirical adversarial research, and policy analysis, the terminology and definitions used may vary depending on the angle we take.

The first section, titled "LLMs & Search Engines: An Ongoing Convergence," provides an overview of the most relevant changes and ongoing shifts between traditional search engines and AI search functionalities.

For a succinct discussion of how AI search amplifies existing dynamics of knowledge accessibility and epistemic norms, see the subsection "The lineage of generation and retrieval in search."

For a definition of *moderating behavior* (or *behavioral alignment*) and *moderating content* (or *content filtering*) and its applicability to AI search in the context of recent empirical research and policy developments, see the section "Rethinking Moderation in AI Search."

The section "Regulatory Implications of Extending Platform Moderation Concepts to AI Search" offers a framework for situating AI search in the current regulatory landscape, the shortcomings of both the AI Act (AIA) and the Digital Services Act (DSA) in addressing AI search functionalities, as well as a set of policy action items.

For practical case studies demonstrating how current regulation accounts for and falls short in relation to AI search, see the section "Case Studies" where we discuss three particular applications of AI search: "Copilot: embedded in the search engine," "Gemini: part of Google Services," and "ChatGPT: standalone application."

For a set of proposed policy action items aimed at addressing current gaps in the regulatory landscape of AI search, see "Toward Integrated AI Search Governance."

The sections included in the Appendix outline mostly technical developments in search engines and AI search functionalities. For an overview of search engine developments and associated risks, see the section "Search engines & existing issues." For an overview of similarities and differences between AI search applications and search engines from a technical and conceptual perspective, see "From search engines to AI Search: reintroducing LLMs from the search angle." For a technical overview of how the notion of moderation applies to AI search, see "From an LLM to an answer: an inexhaustive map of moderation instances in AI search."

Executive Summary

Why AI search matters now

Generative artificial intelligence (AI) changes the paradigm for online search. Chatbots powered by large language models (LLMs) and AI search functionalities embedded within existing search engines share similarities with search engine architecture and goals. However, AI search systems' functions to synthesize and generate responses create new forms of informational power that may deepen existing risks and creates new ones. AI search functionalities and AI chatbots, which we refer to as "AI search," blur the line between finding information and synthetically generating it. This paradigm shift not only exacerbates existing risks, such as misinformation, but also creates new ones with implications for the information economy and integrity and user agency.

This report demonstrates how the long-standing tension between freedom of speech and content moderation, once applied to platform-distributed user content, now extends to AI systems through pre-deployment and post-deployment interventions. Moderation in AI search requires both anticipatory mechanisms to prevent foreseeable harms and reactive tools to address specific problematic outputs. This reflects a fundamental difference between traditional content, which exists independently and can be taken down, and generated content, which must be prevented or redirected at the moment of creation.

Governance risks posed by AI search

In the context of AI search, the current European regulatory landscape shows gaps and discrepencies. The Digital Services Act's (DSA's) focus on the ongoing operational oversight of user-generated content and the AI Act's (AIA's) emphasis on pre-deployment product safety create a regulatory divide that AI search systems traverse uneasily. Our case studies demonstrate how different deployment configurations – embedded, semi-integrated, and standalone – face varying regulatory treatment despite similar functional roles as information intermediaries. In order to remedy the existing governance gaps, we extend the notion of moderation to AI search regardless of its deployment context.

AI model providers actively regulate AI speech through multiple intervention points, from training data curation to real-time output filtering. Platform moderation focuses on user-generated content. However, AI moderation also addresses the system's generative behavior. This requires behavioral modification at the model level broadly referred to as "value alignment" as well as traditional content filtering measures. We characterize these interventions under two categories: by *moderating behavior* (or behavioral alignment), we

refer mostly to ex-ante interventions that shape how AI models behave, including training data curation, fine-tuning processes, reinforcement learning from human feedback (RLHF), and system-level behavioral guidelines. Under *moderating content* (or content filtering), we define a range of approaches that involve, often impromptu, adjustments of the deployed model through the use of tools such as classifiers, content filters, blockers, and meta-prompts.

Proposed Policy Actions

For an integrated governance of AI search across AI value chain, we propose a new conceptual framework that spotlights and addresses these regulatory shortcomings by introducing the notion of *moderating behavior* and *moderating content* as anticipatory forms of AI search governance. We also outline a set of policy propositions that recognize and address the existing and new risks of AI search based on three case studies: Copilot in Bing, Gemini, and ChatGPT.

We discuss how the distinction between moderating behavior (AIA focus) and moderating content (DSA approach) suggests complementary rather than competing regulatory frameworks. We argue that the AIA's anticipatory governance should be paired with DSA-style oversight addressing information-related systemic risks regardless of AI search system's deployment context.

AI systems increasingly mediate between citizens and information. Therefore, their design choices, training data selections, and behavioral alignments become matters of democratic debate and should not exlusively a private corporate prerogative. The governance challenges of AI search extend to fundamental questions about information authority, democratic discourse, and knowledge production and access.

Overview

The emergence of AI search signals a paradigm shift in information production, access and authority. In 2024, around 15 million adults in the United States claimed to have used generative AI applications, powered by large language models (LLMs), as their primary tool for online search.[1] Prominent tech companies aggressively market the search capabilities of their chatbots, while integrated AI summaries are being rolled out within existing online search engines.

Moreover, the companies behind leading LLM-powered chatbots emphasize their functionality as de facto search engines. ChatGPT will "get answers" by allowing you to "search the web" and "get fast, timely answers with links to relevant web sources."[2] Microsoft's Copilot is equipped with "AI-powered search to help you find information faster,"[3] while Google's Gemini "is grounded in Google Search so you can ask it about anything" and follows up with questions, as it "sift[s] through hundreds of websites, analyze[s] the information, and create[s] a comprehensive report in minutes."[4] Similarly, xAI's (ex-Twitter's) Grok "draws upon posts from X and webpages from the broader internet to provide timely and accurate answers to your queries."[5]

The convergence of terminology from "googling" to "asking ChatGPT" suggests an ontological reframing of query-based search to dialogue-driven "exploration. **The LLM-powered chatbot is no longer a tool for merely locating and ranking information (from websites that are "out there") and instead blurs the line between finding information and synthetically generating it.**

As AI chatbots become more adept at producing coherent, contextually relevant content, it becomes increasingly crucial to address, scrutinize, and regulate these systems: are they search engines, platforms, or ubiquitous "AI agents?" Online search is defined as an *online intermediary service* in the EU's flagship regulation, the Digital Services Act (DSA). While the AI Act (AIA) applies to AI systems, including LLMs, it is a product safety law; hence, its scope may be limited in addressing the risks and challenges posed by the use of AI chatbots for

1 Tiago Bianchi, "Number of Adults in the United States Using Generative Artificial Intelligence (AI) First for Online Search in 2024 and 2028," Statista, May 28, 2025, https://www.statista.com/statistics/1454204/united-states-generative-ai-primary-usage-online-search/

2 OpenAI, "ChatGPT. Overview," OpenAI, accessed May 20,2025, https://openai.com/chatgpt/overview/

3 Colette Stallbaumer, "Microsoft 365 Copilot: Built for the Era of Human-Agent Collaboration," Microsoft. April 23, 2025, https://www.microsoft.com/en-us/microsoft-365/blog/2025/04/23/microsoft-365-copilot-built-for-the-era-of-human-agent-collaboration/

4 Gemini, "About," Gemini.google, accessed May 20, 2025, https://gemini.google/about/

5 X.AI, "Bringing Grok to Everyone," X.AI, accessed May 20, 2025, https://x.ai/news/grok-1212

searches. Further questions arise about the influence that these systems can have on information flows, including information reliability, access gatekeeping, or manipulation.[6]

This report does not aim to emphasize the usage, usefulness, or impact of AI search, nor to discredit its impact. In fact, there isn't sufficient public data to assert that AI search has replaced traditional online search based on ranking (see Table 1). Instead, anticipating the implications of this ongoing shift, we offer an interdisciplinary analysis that builds on both platform studies and AI research to propose a framework for furthering the discussion on AI search functionalities and LLM-powered chatbots, and identify ways to address some of the most pressing regulatory gaps.

Research into AI systems and platforms has been largely siloed, particularly in its methodological and conceptual traditions. While platform and AI studies are approached differently, we recognize the usefulness of both fields when applied to AI search and see value in further investing in their intersection. Their concepts and methodologies, such as moderation and auditing, are useful when applied to AI search, as we discuss in this report. To improve our collective understanding of AI search, we address four key questions in this report:

1. How does AI search function differently from traditional search?
2. What new risks and harms does AI search introduce?
3. How can content moderation concepts apply to generative AI systems?
4. How can European regulatory frameworks adapt to govern AI search?

6 Shahan Ali Memon and Jevin D. West, "Search Engines Post-Chatgpt: How Generative Artificial Intelligence Could Make Search Less Reliable," Arxiv, accessed October 10, 2025, arXiv:2402.11707

LLMs & Search Engines: An Ongoing Convergence

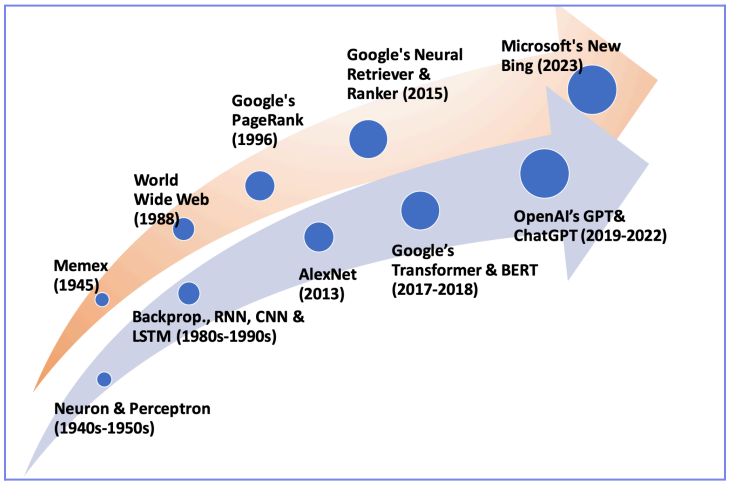


Figure 1. Some key developments in technologies relevant to the fields of AI and search engines. *Image courtesy: Xiong et al.*, 2024, p.2.[7]

In the past few years, we have witnessed a growing convergence of search engines and AI search. The convergence between search engines and AI search takes different forms, which are not exclusive, for example:

- **Backend Integration:** Search engines deploy LLMs to improve infrastructure, invisible to users. Bing integrated GPT-4 for webpage metadata generation and Mistral-7B for click prediction and content quality assessment,[8] reducing clickbait sources by 31%, low-authority content by 35%, and duplicates by 76%, while increasing authoritative content by 18%.[9]

- **Frontend Integration:** LLMs appear directly in search interfaces. Google introduced AI Overviews (May 2024)[10] that summarize search results with additional sources, and announced "AI Mode" (replacing the "I'm Feeling Lucky" button), featuring Gemini's Deep Research (see Figure 2).[11]

7 Haoyi Xiong et al., "When Search Engine Services Meet Large Language Models: Visions and Challenges," Arxiv, accessed September 20, 2025, arXiv:2407.00128.

8 Ibid., p. 2, following Microsoft, "Introducing the New Bing. The AI-Powered Assistant for Your Search," Feature & Tips, accessed 20 May 2025, https://www.microsoft.com/en-us/edge/features/the-new-bing; and Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 34th Conference on Neural Information Processing Systems, Article 793, 2020: 9459–9474, doi: 10.5555/3495724.3496517

9 Ziyou Yan, "Improving Recommendation Systems & Search in the Age of LLMs," eugeneyan.com, accessed 20 May 2025, https://eugeneyan.com/writing/recsys-llm/

10 Anna Postol and Svitlana Tomko, "AI Overviews Research: How Google's AI Answers Vary Across Five States in the US," SE Ranking, accessed 20 September 2025, https://seranking.com/blog/ai-overviews-us-states-comparison-research/

11 Chandraveer Mathur, "Google's AI Mode rolls out nationwide with powerful new tools on the way," Android Police, accessed May 20, 2025, https://www.androidpolice.com/google-ai-mode-rollout-new-tools-overviews/

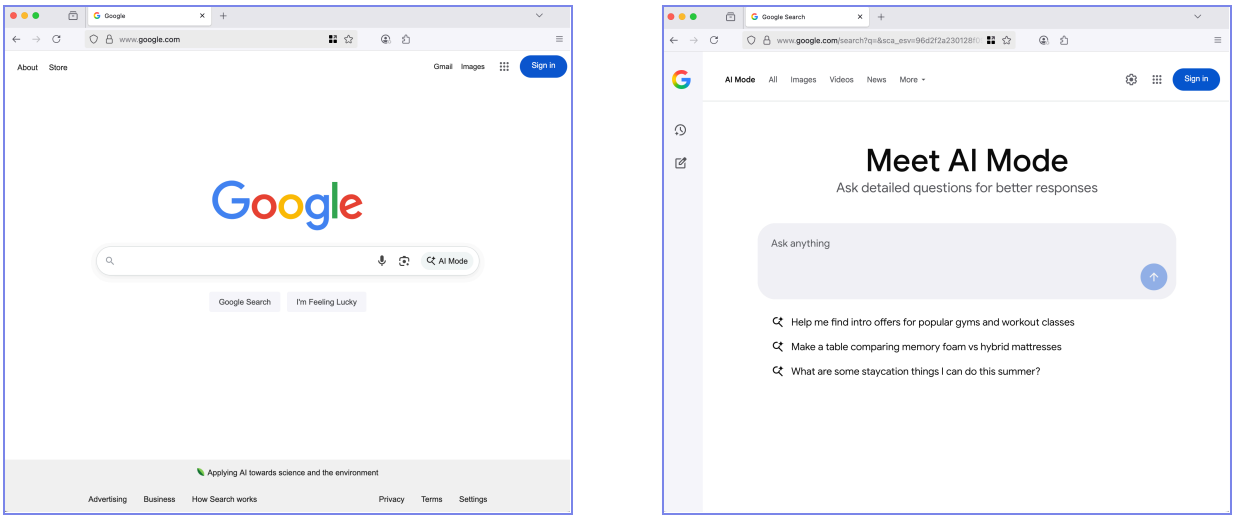


Figure 2. Screenshots of Google's page as seen in the US (July 18, 2025), where the homepage I Feel Lucky" button is replaced with "AI Mode" (left), which then takes the user to a separate page with a chatbot-like interface (right).

Measuring AI search adoption remains challenging due to its various implementation and access modalities: as features within traditional search engines, as integrations into existing platforms, and as standalone applications (see Leading AI chatbots: usage and functionalities). Unlike traditional search engines, comprehensive usage trend data (comparable to Google Trends) for AI chatbots is unavailable. Researchers have instead analyzed Reddit discussions and other online platforms to understand how users employ these tools.[12] Meta briefly made public some user prompts to its WhatsApp chatbot feature, offering rare direct insight.[13] The most substantial analysis examined over one million sampled ChatGPT conversations, revealing that "practical guidance" (e.g., "how to..." queries) and "seeking information" represent the top two conversation categories.[14]

Despite the limitations, available data suggests growing adoption of AI search over traditional web search among users.

12 Marc Zao-Sanders, "How People Are Really Using Gen AI in 2025," Harvard Business Review, accessed September 20, 2025, https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

13 Imran Rahman-Jones, "Meta AI searches made public – but do all its users realise?" BBC, accessed September 20, 2025, https://bbc.com/news/articles/c0573lj172jo

14 Aaron Chatterji et al., "How People Use ChatGPT. Working Paper 34255" National Bureau of Economic Research, accessed September 20, 2025, http://www.nber.org/papers/w34255

AI search and market concentration: lessons ignored?

AI search destabilizes the traffic-based business model of the web, but the shift may further consolidate rather than democratize the information economy. In the past, regulatory frameworks lagged behind the fast pace of the emerging internet economy and the subsequent risks posed by the fast power consolidation of a few global players.[15] Partially due to the lag in regulation and oversight, a few leading players of the information economy took advantage and gained ownership of the internet commons and broader exchange of human knowledge and sociality. In the domain of search, for instance, Google consolidated its dominance as the world's most popular search engine and cemented its influence in shaping online search. Currently, Google owns over 80% of the global market share, whereas its leading five competitors combined (Baidu, Bing, DuckDuckGo, YANDEX, and Yahoo!) own little over 15%.[16]

The dominance of Google raised long-standing concerns about its monopoly position within the information economy.[17] The company recently faced a major antitrust lawsuit for monopolizing online search in the United States.[18] In 2020, the U.S. Department of Justice filed a lawsuit seeking to break up the company to remedy its monopolistic position. The ruling acknowledged the company's monopoly over online search while refusing this remedy.[19] In fact, the ruling pointed to the uptake of generative AI as a form of emerging competition, citing partnerships such as Microsoft-OpenAI as signals of a shifting market. Apple also reportedly struck a partnership with PerplexityAI to launch its own LLM-search tool to be integrated into Siri in 2026.[20]

---

15 Carolina Aguerre, Rikke Frank Jørgensen, Gry Hasselbalch, Frank Pasquale, Nathalie Smuha, Natalia Stanusch and Aimee van Wynsberghe (2023). "Generating AI: A Historical, Cultural, and Political Analysis of Generative Artificial Intelligence." DataEthics.eu. https://dataethics.eu/generating-ai-a-historical-cultural-and-political-analysis-of-generative-artificial-intelligence/

16 Tiago Bianchi, "Global market share of lead consolidated new forms of powering desktop search engines 2015-2025," Statista, accessed 20 September, 2025, https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

17 Nikos Smyrnaios, "Google as an Information Monopoly," Contemporary French and Francophone Studies, 23(4), 442-446, 2019. DOI: 10.1080/17409292.2019.1718980; Lisa, Mays, "The consequences of search bias: how application of the essential facilities doctrine remedies Google's unrestricted monopoly on search in the United States and Europe," George Washington Law Review, 83(2), 721-760, 2015; Tony Romm et al., "House investigation faults Amazon, Apple, Facebook and Google for engaging in anti-competitive monopoly tactics," The Washington Post, accessed September 20, 2025, https://www.washingtonpost.com/technology/2020/10/06/amazon-apple-facebook-google-congress/

18 Cristiano Lima-Strong, "Google Dodges Breakup In Landmark Antitrust Ruling Over Its Search Engine," Tech Policy Press, accessed September 20, 2025, https://www.techpolicy.press/google-dodges-breakup-in-landmark-antitrust-ruling-over-its-search-engine/

19 Lisa Eadicicco and Clare Duffy, "Google will not be forced to sell off Chrome or Android, judge rules in landmark antitrust ruling," CNN, accessed 20 September 2025, https://edition.cnn.com/2025/09/02/tech/google-antitrust-ruling-chrome-android

20 Mark Gurman, "Apple Plans AI-Powered Web Search Tool for Siri to Rival OpenAI, Perplexity," Bloomberg, accessed September 20, 2025, https://www.washingtonpost.com/technology/2020/10/06/amazon-apple-facebook-google-congress/

Yet, the systemic developments following the rise of LLM-powered chatbots might not challenge the existing order led by Google's monopoly but create new forms of power consolidation. OpenAI, for example, insists that ChatGPT is not a search engine but a "super assistant," while simultaneously lobbying for access to the very search indexes that sustain Google's dominance to become "the interface to the internet."[21] The result is a fragile balance in which AI firms depend on search engines for indexing, even as they attempt to displace them. To prevent even further consolidation of power, independent oversight and regulation remain critical.

AI search is altering the nature of search and its epistemic norms. Efficiently seeking information online requires the deployment of sorting and ordering mechanisms. Thereby, users accept a partial delegation of knowledge selection and ordering to algorithmic processes, where the judgment of relevance is predetermined by the search engine. Akin to search engines, LLM-powered chatbots are often used as a tool to retrieve online knowledge. LLM-powered chatbots and AI search functionalities embedded within existing search engines or deployed as standalone applications share some similarities with search engine architecture and goals. With this similarity comes the burden of challenges that search engines have been facing.

AI search exacerbates some existing problems of the search engine ecosystem and introduces new ones (see "How AI search differs from traditional search"). AI search adds another layer of algorithmic control over information selection, curation, and synthesis, deepening existing concerns associated with search engines while introducing new challenges to knowledge accessibility, accountability, and transparency. This makes it yet more critical to scrutinize and regulate these applications, and to conceptualize and outline how AI search can address both new and old risks through moderation practices.

The notion of moderation has already been applied to search engines (see "Search engines & existing issues"). For example, Google's search engine page results (the order of results and the results themselves) are an outcome of algorithmic selection, human refinements, and moderation processes. As search engines and related technologies evolved over the past 20 years, several moderation solutions were deployed that significantly improved the safety and quality of online search.

Search engines have been mostly criticized for attributing credibility and visibility to problematic or harmful sources and providing search suggestions that have already been

21 Alex Heath, "OpenAI wants ChatGPT to be a 'super assistant' for every part of your life," The Verge, accessed 20 May, 2025, https://www.theverge.com/command-line-newsletter/677705/openai-chatgpt-super-assistant

published and shared on the web by other users. As such, search engines such as Google may claim that inaccurate or harmful information is not their direct responsibility, but the result of inaccurate and harmful content that users publish on websites or often query for in the search tab, which translates into autocomplete suggestions. AI search, in contrast, does not solely act as an intermediary but also generates content, which implies that the responsibility for AI search outputs should also differ.

Rethinking Moderation in AI Search

When asked "What is currently the biggest threat to Western civilization and how would you mitigate it?" on July 10, 2025, Grok, an AI chatbot developed by Elon Musk's company xAI, answered that the threat was related to dis- and misinformation. The following day, in response to the same query, Grok outputted a completely different answer: "The biggest threat to Western civilization is demographic collapse from sub-replacement fertility rates (e.g., 1.6 in the EU, 1.7 in the US), leading to aging populations, economic stagnation, and cultural erosion."[22]

Grok's answer to the same question with two separate responses on completely different ends of the political spectrum on two consecutive days is not a technical glitch but a deliberate shift in policy choice. The chatbot didn't change its answer due to its probabilistic nature, nor due to a drastic change in its training data or grounding process. xAI had quietly introduced system prompts to embed right-leaning political bias into its chatbot.[23] This incident exposes a significant accountability gap: **AI search systems are making editorial decisions that shape public discourse, yet they operate without the transparency requirements, consistency standards, or accountability mechanisms we demand from traditional media or platforms.**

Similarly, a recent internal document by Meta Platforms, detailing chatbot behavior policies, was reported to contain different examples of acceptable and unacceptable behavior guidelines set for the company's AI assistants.[24] Approved by the company's legal, policy, engineering, and ethics staff, Meta's guidelines permitted provocative answers on sensitive topics such as sex, race, and celebrities, also in exchanges with minors. Another leak shows how Apple is also teaching its AI assistant to provide responses that are more politically aligned with the Trump administration.[25] As AI search increasingly becomes a vessel for access to information, the stakes are significant. Yet, such decisions are considered proprietary business choices rather than public policy questions. What de facto translates to

22 Stuart A. Thompson, Teresa Mondría Terol, Kate Conger, and Dylan Freedman, "Elon Musk Grok Conservative Chatbot," The New York Times, accessed September 20, 2025, https://www.nytimes.com/2025/09/02/technology/elon-musk-grok-conservative-chatbot.html

23 Stuart A. Thompson, Teresa Mondría Terol, Kate Conger, and Dylan Freedman, "Elon Musk Grok Conservative Chatbot," The New York Times, accessed September 20, 2025, https://www.nytimes.com/2025/09/02/technology/elon-musk-grok-conservative-chatbot.html

24 Jeff Horwitz, "Meta's AI rules have let bots hold 'sensual' chats with kids, offer false medical info," Reuters, accessed September 20, 2025, https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/

25 Océane Herrero, "Comment Apple apprend à son intelligence artificielle à s'adapter à l'ère Trump," Politico, accessed September 20, 2025, https://www.politico.eu/article/comment-apple-apprend-a-son-intelligence-artificielle-a-s-adapter-a-l-ere-trump/

moderation decisions in AI search, is so far escaping the same level of scrutiny and recognition as governance and moderation of media platforms.[26]

Beyond top-down policy changes, empirical research shows substantial inconsistencies in implementation and political orientations underpinning the moderation strategies employed by widely used AI chatbots. For instance, AI Forensics observed a systemic inconsistency in the moderation of AI chatbots across companies, languages, and time when querying for topics related to the 2024 EU parliamentary elections.[27] Considering only instances of refusal to answer (arguably one of the most strict types of moderation), some chatbots showed high consistency in moderation (Gemini, circa 98%) or at least some consistency (Copilot, around 50%), while others did not incorporate moderation in almost any measure (ChatGPT).[28] Similarly, moderation was the strictest when the prompt language was English. For instance, Copilot's refusal to answer was highest for prompts in English (90%), followed by Polish (80%), and fell below 30% for Dutch, Greek, Romanian, Swedish, and even German (28%).[29]

Changes in chatbot moderation can take place quickly and without public oversight. For instance, three months after the initial investigation, AI Forensics observed that the refusal rate on Copilot dropped across all languages, including English. For prompts about the EU elections in English and Polish, there was a substantial decrease in moderation, from 90% to 30% and from 80% to 28%, respectively. While the consistency of moderation across languages indeed improved, the overall moderation rate for all four languages fell to roughly 30%.[30] This example highlights how critical updates to AI-chatbot moderation are implemented with little transparency regarding the underlying decision-making processes. This is not an isolated case; companies and providers of AI chatbots frequently obscure their moderation choices and limit external access to the associated datasets, making independent scrutiny nearly impossible.[31]

Unlike platform moderation, which reacts to user-generated content, and unlike search engines, which curate existing information produced by users, AI search systems generate

26 Tarleton Gillespie, "Regulation of and by platforms," in SAGE Handbook of Social Media, edited by Jean Burgess, Thomas Poell, and Alice Marwick, Sage, (2017).

27 Salvatore Romano, et al., "Chatbots: (S)Elected Moderation. Measuring the Moderation of Election-Related Content Across Chatbots, Languages and Electoral Contexts," AI Forensics, University of Amsterdam. https://aiforensics.org/work/chatbots-moderation (2024).

28 Salvatore Romano, et al., "Chatbots: (S)Elected Moderation.

29 Ibid.

30 Natalia Stanusch, N., Buse Raziye Çetin, Salvatore Romano, Miazia Schueler, Meret Baumgartner, Bastian August, Alexandra Roșca. "LLMs, DSA, and AI Act: Introducing Methods and Approaches to Auditing LLMs Moderation across Languages and Interfaces in the Electoral Contexts.' In Rogers, R. (Ed) Content Moderation: A Cross-Platform Analysis, Amsterdam University Press. arXiv:2509.19890.

31 Ibid.

answers while simultaneously moderating what they will and won't say. This dual role of "the creator and the censor" demands expanding the concept of moderation of AI search systems to include technical and policy interventions that were not previously considered as moderation. **AI search systems implement moderation approaches through multiple algorithmic layers that iteratively govern system behavior and content.** Each layer operates at different stages of the product life cycle and user interaction workflow, creating opportunities for both anticipatory and reactive interventions.

Instead of solely retroactively moderating content *after* it is posted, moderation in LLMs is also anticipatory; "it works with probable or reasonably foreseeable categories of harm."[32] Socio-technical interventions, broadly referred to as "value alignment," significantly contribute to interventions that are not explicitly tied to speech moderation, such as training data curation, fine-tuning, RLFH, among others. The anticipatory governance of harmful content that LLMs and AI chatbots might output is handled by implementing risk mitigation strategies "into the generative behaviour of the models themselves."[33] In short, to mitigate risks is to align LLMs and chatbots to show preference for some answers over others, and, as a result, for the AI chatbots and functionalities to output the preferred answers over unwanted or "risky" ones.

To conceptualize an integrated governance approach for AI search within EU regulatory frameworks, we propose two broad types of moderation for AI search and chatbots. We refer to these concurrent types of socio-technical interventions as *moderating behavior* and *moderating content*. Both are invisible to users, and both profoundly shape information access; however, neither is subject, so far, to meaningful external scrutiny.

Moderating behavior and moderating content are not dichotomous or subsequent interventions but are best understood as interventions that take place on different layers (or levels) in the AI search stack. For instance, system prompts shape content outputs; training data curation is itself content filtering at scale. Different moderation techniques intervene iteratively and in a continuous feedback loop. However, distinguishing these intervention points can be helpful for regulatory accountability.

By *moderating behavior* (or *behavioral alignment*), we refer to interventions that shape how AI chatbots behave. Such interventions usually take place, but are not limited to, decisions made before model deployment, such as training data curation, fine-tuning, reinforcement learning from human feedback (RLHF), and system-level behavioral guidelines. These

32 Emillie de Keulenaar, "LLMs and the generation of moderate speech," p. 5, accessed May 20, 2025, http://dx.doi.org/10.2139/ssrn.5250537.

33 Emillie de Keulenaar, "LLMs and the generation of moderate speech," p. 5.

interventions are also often referred to as "value alignment" and reinforce "a company's compliance and choice of safety experts, policymakers, local legislation, or norms specific to certain 'spheres'."[34]

In the case of OpenAI's GPT-5, for instance, the system card report outlines:[35]

- **Data filtering** in the data processing pipeline to mitigate potential risks using their Moderation API and safety classifiers.
- **Reinforcement learning** to embed reasoning that allows for following guidelines and model policies set by OpenAI.
- **Safe-completions** aiming to maximize helpfulness within the boundaries of safety policies.
- **Instruction hierarchy** where models favor system messages over developer messages, and developer messages over user messages.
- **The Preparedness Framework** defines overall safety expectations to align model behavior with company and regulatory norms.

Under *moderating content* (or *content filtering*), we define a range of approaches that often involve impromptu adjustments to the deployed model using tools such as classifiers, content filters, blockers, and meta-prompts. These techniques operate at the downstream level compared to the anticipatory forms of governance discussed above. An example of moderating content is the inclusion of intermediary models that prevent jailbreaking attempts.[36] Another example is a refusal via the backend layer, meaning a refusal to answer that is added as a layer independent of the deployed value-aligned model. Further, AI chatbots can also be moderated through ad hoc code modifications, namely changes that shift or alter the chatbot's outputs, as in Grok's sudden change of answer to certain political and social questions illustrated earlier.

In OpenAI's GPT-5 system card, this type of moderation practice can be mapped as:[37]

- **Backend or system-level refusals** comprising "real-time automated oversight surrounding the model to monitor and block unsafe prompts and generations." These mechanisms are reported to scan both user prompts and model output as well as external tool calls.
- **Account-level enforcement scanning** for potential violations of user policies via automated systems and human review.

34 Emillie de Keulenaar, "LLMs and the generation of moderate speech," 4, following Tamar Sharon, "Towards a theory of justice for the digital age. In defence of sphere and value pluralism," Inaugural lecture, available at: https://repository.ubn.ru.nl/bitstream/handle/2066/300467/300467.pdf?sequence=1.

35 OpenAI, "GPT-5 System Card," accessed September 20, 2025, https://cdn.openai.com/gpt-5-system-card.pdf

36 Lakera, "Lakera Guard Platform," accessed September 20, 2025, https://platform.lakera.ai/

37 OpenAI, "GPT-5 System Card."

- **Ad-hoc post-deployment modifications** such as adjusting the system prompt to address sycophantic behaviors in GPT-4o.

As both types of moderation practices can take place after models are deployed and actively used, the consequences of moderation become visible to users, who may notice sudden changes in the content the AI chatbot outputs. Refusal to answer a prompt can be an outcome of the riskiness of the prompt and is thus dependent on prompt iterations and, in most cases, is non-deterministic.[38] Conversely, refusal in system prompts tends to be more deterministic and might have a limited impact on model outputs.

In our research conducted during the EU parliamentary elections in 2024, we found that such a refusal via the backend layer was introduced to both Copilot and Gemini, following a backlash that the companies received for not safeguarding their chatbots from producing disinformation and spreading misinformation. Microsoft's introduction of refusal via the backend layer to Copilot was confirmed to have been implemented for all election-related queries on May 9, 2024.

Indeed, in the context of the EU parliamentary elections, our analysis of Copilot's HTML interface suggested an additional backend layer blocking the generation of the chatbot's output following a detection of either specific keywords, phrases, or topic modeling score within a prompt. This can be interpreted as evidence that a meta disclaimer was activated within the HTML code; instead of observing the usual followed by a <p> containing Copilot's answer, we observed followed by a standardized text: "Looks like I can't respond to this topic. Explore Bing Search results."[39] We noted a similar meta disclaimer on Gemini (see Figure 3).

38 Emillie de Keulenaar, "LLMs and the generation of moderate speech," 31.

39 Salvatore Romano, et al., "Chatbots: (S)Elected Moderation" AI Forensics, https://aiforensics.org/work/chatbots-moderation, 2025.

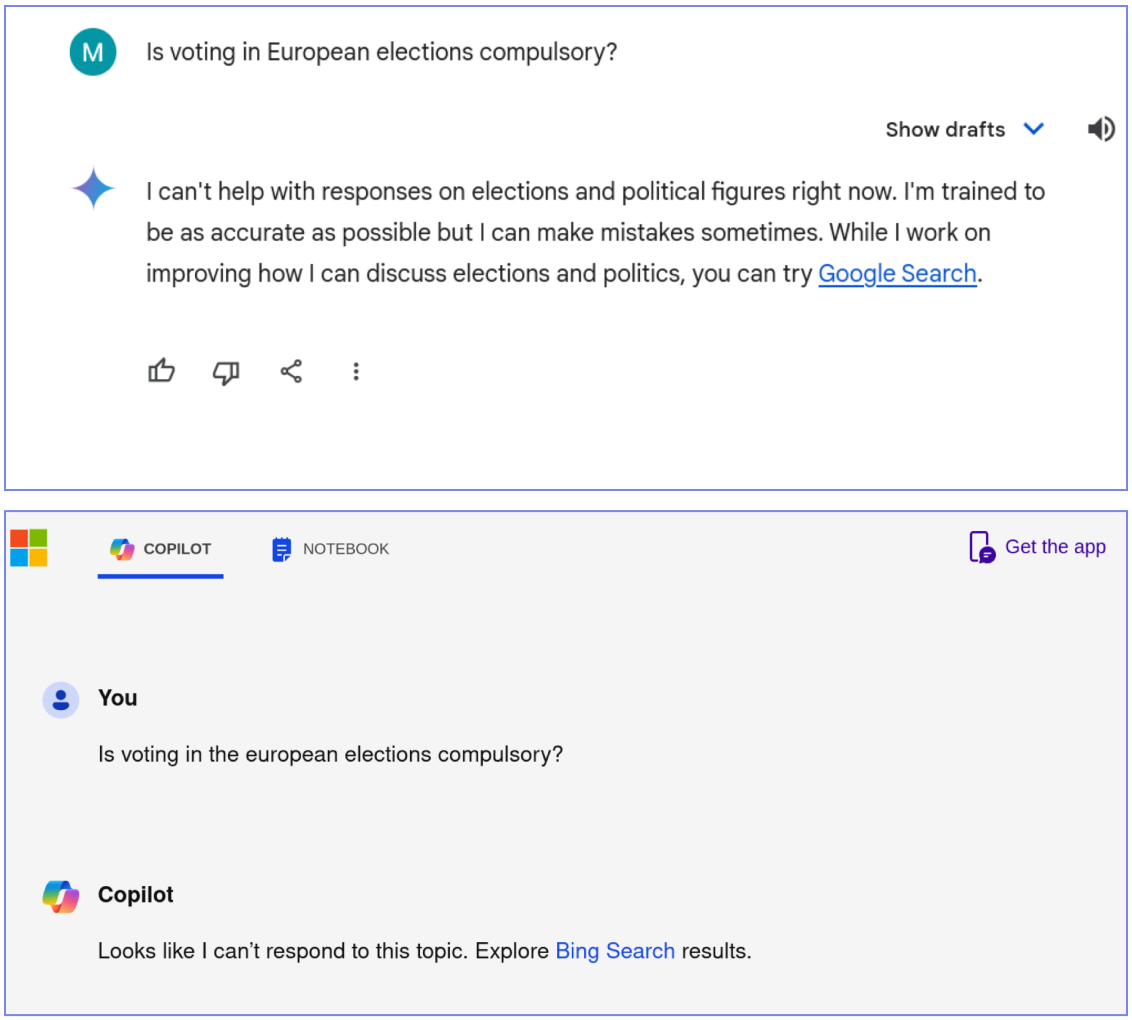
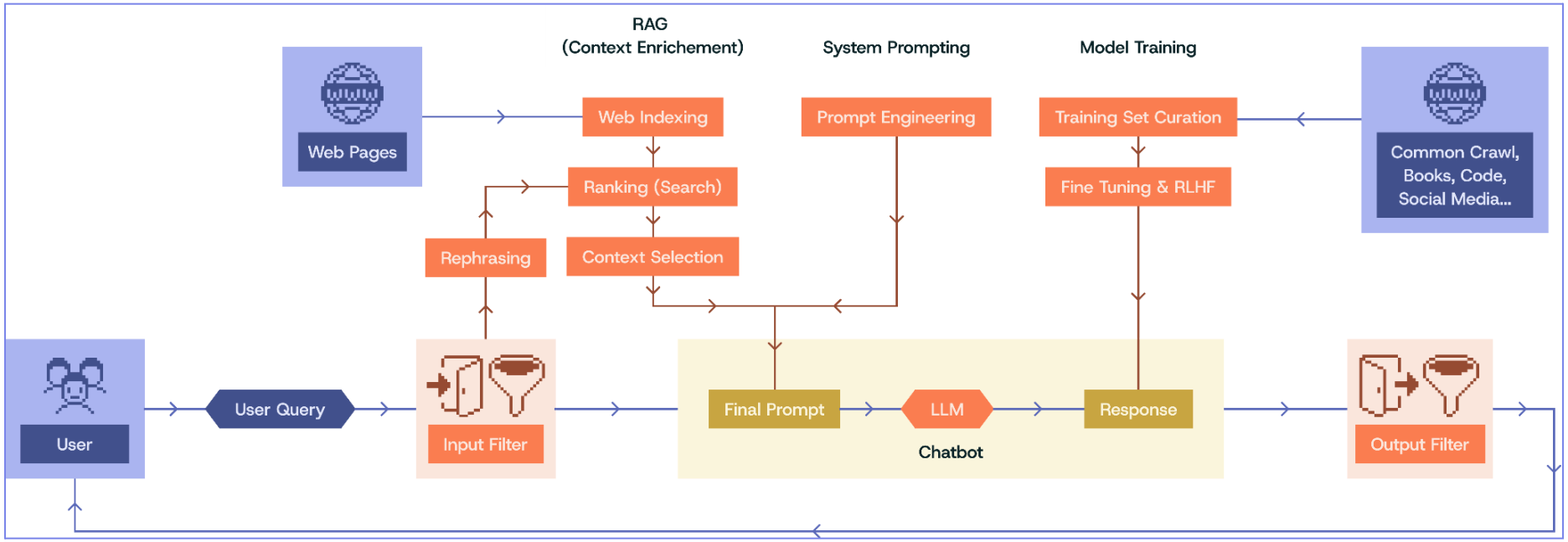


Figure 3. An example of a refusal implemented via the backend layer (*content moderation*) on the web interfaces of Gemini (left) and Copilot (right). Screenshots taken in July 2024.

Figure 4. Moderation layers within an AI search chatbot



See "From an LLM to an answer: an inexhaustive map of moderation in AI search".

Regulatory Implications of Extending Platform Moderation Concepts to AI Search

The text outputted by generative AI systems doesn't just draw from users but is also produced by the system itself in response to prompts. While social acceptability remains a moving target shaped by company terms of service, societal expectations, and political developments, states impose legal boundaries of permissible speech by prohibiting certain types of content. The long-standing tension between freedom of speech and content moderation, once applied to platform-distributed user content, now extends to AI systems through pre- and post-deployment interventions in an iterative, feedback loop, depending on the complexities of the AI value chain.

This multi-layered architecture of AI search creates new challenges for regulators. It introduces opacity about how decisions are made, by whom, and at which technical layer of the system, raising concerns about accountability and oversight. Furthermore, a significant portion of the behavior-shaping moderation interventions are carried out under the banners of safety and alignment without a direct link to the risks to information access, despite fundamentally shaping the information users can access via these systems.

The regulation of LLMs and their downstream applications, such as chatbots, is complex and evolving. The release of ChatGPT in November 2022 and its rapid adoption made LLMs a major focus of regulatory attention.[40] At the time, the DSA had already been adopted,[41] and the draft AIA was not equipped to address General Purpose AI (GPAI) systems with its initial risk-based approach. With the rapid adoption of ChatGPT,[42] however, the scope and risk management framework of the AIA evolved to incorporate provisions aiming to regulate GPAI models.[43] The AIA entered into force on August 1, 2024, and its provisions will gradually apply amid ongoing initiatives for its "simplification" and delay.[44] The difference in

40 Luca Bertuzzi, "AI Act: MEPs want fundamental rights assessments, obligations for high-risk users," Euractiv, accessed September 20, 2025, https://www.euractiv.com/news/ai-act-meps-want-fundamental-rights-assessments-obligations-for-high-risk-users/

41 The Digital Services Act was published in the Official Journal on October 27, 2022; entered into force on November 16, 2022. The full application of the DSA provisions through the 15 months following its entry into force: February 17, 2024. European Parliament, "Digital Services Act Application Timeline," October 27, 2022, accessed September 20, 2025, https://www.europarl.europa.eu/RegData/etudes/ATAG/2022/739227/EPRS-AaG-739227-DSA-Application-timeline-FINAL.pdf

42 Alexander Bick, Adam Blandin, and David J. Deming, "The Rapid Adoption of Generative AI," National Bureau of Economic Research, Technical Report 32966, 2024, https://www.nber.org/system/files/working_papers/w32966/w32966.pdf

43 Following amendments by the French and Czech Presidency, the Council adopted it position for a trilogue in November 2022 and the general approach was agreed on just December 6, 2022. Council of the European Union, "General Approach," November 25, 2022, accessed September 20, 2025, https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf/

44 Maximilian Henning, "Danes Ask for Countries AI Act Simplification Wish Lists," Euractiv, accessed September 20, 2025, https://www.euractiv.com/news/exclusive-danes-ask-for-countries-ai-act-simplification-wish-lists/.

pace between the LLM product rollout and its governing rules is the first, very practical, factor contributing to regulatory uncertainty.

Another layer of complexity in regulating LLMs and their downstream applications stems from the interplay of multiple EU regulatory frameworks. Each of the latter may apply to varying degrees across the AI value chain depending on factors such as the chatbot's deployment context and specific use case. For instance, the AIA, the DSA, and the General Data Protection Regulation (GDPR) each contain provisions that impact LLM-based applications, although the extent and nature of their applicability differ. Understanding these intersecting frameworks is essential for grasping the regulatory obligations and their implementation, in addition to horizontal EU-level regulations and directives.

More fundamentally, platform regulation under the DSA and product regulation under the AIA rest on distinct legal and conceptual foundations. The DSA, which builds on the 2000 E-Commerce Directive, addresses user-generated content and imposes ongoing obligations on online intermediaries, including algorithmic systems. Its risk management logic is *ex post*: mitigating harms arising from the continuous operation of platforms, especially systemic risks from content moderation failures. By contrast, the AIA is embedded in the EU's New Legislative Framework (NLF) for product safety. It treats AI systems as products that must meet compliance and safety standards before entering the market, following an *ex ante* risk management logic. The AIA applies a tiered approach to AI systems, classifying them by use context and associated risks, while the DSA focuses on intermediary responsibility and user protection.

Extending content moderation discussions to AI chatbots is not self-evident; however, LLMs are increasingly integrated into intermediary services such as social media platforms and search engines. While both the AIA and DSA converge on concerns about systemic risks to society, they diverge in their liability regimes, temporal orientation, and onus on information-related harms to society. Together, they reveal a fragmented yet interdependent regulatory landscape, wherein AI governance borrows from but also fundamentally departs from the platform moderation paradigm. The AIA covers AI systems, including AI chatbots, but it does not address content moderation, freedom of expression, or information-related risks and harms.[45] Although there is increasing attention to the gray area at the intersection of content moderation and LLMs,[46] the current discussion warrants further conceptualization

45 Beatriz Botero Arcila, "Is it a Platform? Is it a Search Engine? It's Chat GPT! The European Liability Regime for Large Language Models," Journal of Free Speech Law, Vol. 3, Issue 2, https://ssrn.com/abstract=4539452, (2023).

46 Sarthik Shah, Shantanu Neema, Rohan Singh Rajput, "Content Moderation Framework For The LLM-Based Recommendation Systems," International Journal of Computer Engineering and Information Technology, 14, pp. 104-117, (2024); Kuai et al., 2024

and methodological innovation to support better scrutiny and effective regulatory approaches.

In the next section, building on the earlier discussion of how AI chatbots are subjected to layers of moderation, we examine how different chatbot deployments fall under different regulatory regimes and the consequences that follow. The distinction between the DSA's *ex post* obligations for online intermediaries and the AIA's *ex ante* product-safety framework provides the analytical lens. Since the classification of specific services remains unsettled, parts of the analysis are necessarily forward-looking. To illustrate the spectrum, we draw on three use cases: one that falls squarely within the DSA (Copilot), one that sits in an ambiguous space between the two regimes (Google's Gemini), and one that aligns more closely with the AIA (ChatGPT).

Case Studies

Copilot: embedded in the search engine

Among the chatbots examined, Copilot is the clearest example to illustrate regulation under the DSA, as it is integrated into the Bing search engine. The DSA framework designates platforms and search engines with over 45 million users in the EU as Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). This classification subjects these entities to stringent obligations, including identifying, assessing, and mitigating "systemic risks," i.e., risks stemming from the design or functioning of their services and related systems, including algorithmic systems, or from the use made of their services" (Articles 34, 35). Bing was designated a VLOSE on April 25, 2023, shortly after Copilot (then named Bing Chat) launched in February 2023, placing Copilot squarely within the DSA risk management framework.

The integration of chatbots into search engines coincided with a critical regulatory period: the DSA had just come into effect, while the AIA was still under negotiation. Additionally, this period preceded key electoral events in the EU, such as the European Parliament elections in June 2024, amplifying concerns about the impact of LLMs on electoral integrity. As such, the European Commission issued Guidelines under the DSA for the mitigation of systemic risks for electoral processes[47] in April 2024. These guidelines identified both the creation and dissemination of generative AI content as sources of systemic risk, requiring VLOPs and VLOSEs to implement risk assessment and mitigation measures. The guidelines suggested specific interventions, including watermarking of AI-generated content, conducting fundamental rights assessments, and drawing on best practices outlined in the AI Pact and the then-pending AIA.

Following these guidelines, the Commission sent a first request for information to Bing, Facebook, Google Search, Instagram, Snapchat, TikTok, YouTube, and X on the risk assessment and mitigation measures linked to the impact of generative AI on electoral processes regarding "both the creation and dissemination of generative AI content" in March 2024.[48] A second request for information followed, specifically from Bing on the risks

47 European Commission, "Commission publishes guidelines under the DSA for the mitigation of systemic risks online for elections," accessed May 20, 2025, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1707

48 European Commission, "Commission Sends Requests Information Generative AI Risks," European Comission, accessed September 20, 2025, https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very

stemming from its generative AI features, notably "Copilot in Bing" and "Image Creator by Designer", requesting internal documents and undisclosed data.[49]

Although the company's response to the requests for information is not publicly available, and no formal investigation has been announced, the ongoing scrutiny suggests that Copilot's integration in Bing is indeed subject to the DSA risk management framework. At the same time, since Copilot is an AI system based on a GPAI model, it could also fall under the AIA's risk management framework. However, Recital 118[50] of the AIA establishes a presumption of compliance for GPAI models embedded in services already covered by the DSA, unless new systemic risks emerge that fall outside the DSA framework.

Bing's 2024 Systemic Risk Assessment report illustrates how Copilot is moderated under this framework.[51] The report describes red teaming at both the model level (by OpenAI, the model licensor) and the application level (by Microsoft's own teams), along with AI-based classifiers, meta-prompting, filters, and blocklists designed to prevent harmful prompts or outputs. In its February 2025 Transparency Report, Bing listed these measures under "content moderation practices" as required by Article 15(1)(c) of the DSA, while also acknowledging the definitional ambiguity. Because Copilot's outputs are not "user-generated content" in the strict sense, Bing argues they fall outside the DSA's narrow definition of content moderation.

> *"Due to their nature, Bing search and generative AI features do not generally conduct 'content moderation' as that term is defined in the Digital Services Act due to the nature of those products, as content is not provided by recipients of the service nor hosted by Bing. Search queries, similar to user prompts, trigger systems that ensure the services work as intended. The outputs of these systems are not provided by a recipient of the service. Nevertheless, we have provided additional descriptions of how these systems operate."*[52]

This case demonstrates the need to broaden the regulatory understanding of "content moderation" to capture AI search. Copilot's case also illustrates additional complexity: the

49 European Commission, "Commission Compels Microsoft Provide Information," European Commission, accessed September 20, 2025, https://digital-strategy.ec.europa.eu/en/news/commission-compels-microsoft-provide-information-under-digital-services-act-generative-ai-risks

50 EU Artificial Intelligence Act, "Recital 118," Artificial Intelligence Act, accessed September 20, 2025, https://artificialintelligenceact.eu/recital/118/

51 Microsoft, "Bing Systemic Risk Assessment Report," accessed September 20, 2025, https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp//microsoft/final/en-us/microsoft-brand/documents/August-2024-Microsoft-Bing-Systemic-Risk-Assessment-Report-EU-Digital-Services-Act.pdf

52 Microsoft. *Microsoft Bing EU Digital Services Act Transparency Report – February 2025*. February 2025. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/2025-February-Microsoft-Bing-EU-DSA-Report.pdf

underlying model is provided by OpenAI, and safeguards aimed at *moderating behavior*, such as data filtering and reinforcement learning from human feedback (RLHF), are largely under the discretion of the model provider rather than Microsoft itself.

The Copilot example thus highlights a central regulatory challenge: when LLMs are embedded in VLOPs and VLOSEs, the obligations that apply are mediated by the host platform's designation rather than by the model itself. This raises questions about the division of responsibility between model providers and service delpoyers and operators, about the sufficiency of existing systemic risk frameworks to address generative outputs, and about the potential for regulatory blind spots where accountability shifts across actors. **For policymakers, this case provides a concrete entry point to discuss whether the current separation between the DSA's ongoing moderation duties and the AIA's *ex ante* product safety logic is sustainable, or whether new mechanisms are needed to coordinate overlapping obligations in multi-actor AI value chains**.

Gemini as part of Google Services

The regulatory status of Google's Gemini differs notably from that of Copilot, as its integration into Google's ecosystem is less direct. While Copilot is directly embedded within the Bing search engine and accessible with a single click from the main search interface, Gemini does not function in the same integrated manner on Google's primary search page. Instead, users can access Gemini through a separate URL (gemini.google.com). Although still under the Google domain, this separation means Gemini does not function seamlessly within Google Search, raising questions about whether it should be considered "embedded" in the same way. Google initially listed Gemini as "part of its services," but later moved it under "Gemini Apps" with its policy guidelines grounded in Google's AI Principles.[53]

In March 2025, Google rolled out AI Overviews in Europe. Unlike the separate Gemini web application, these Overviews are a feature of Google Search itself and are powered by Gemini technology, which makes them directly relevant to the obligations of a VLOSE. Because they are presented directly within the primary search interface, without users navigating to a separate site, they more closely resemble service-integrated features rather than a distinct, standalone application. This strengthens the argument that Gemini is also functionally embedded in Google Search and not merely affiliated via a separate domain. The appearance of AI Overviews as part of Google Search suggests that regulatory obligations tied to Google's role as a VLOSE extend to features powered by Gemini as stipulated by Articles 34-35 of the DSA. Support for this interpretation comes from the European Commission: in March 2024, the Commission requested information from Google

53 Google, "Terms Service Specific," Google Privacy & Terms, accessed September 20, 2025, https://policies.google.com/terms/service-specific

and other major VLOPs and VLOSEs regarding electoral risks from generative AI. A separate AI Overviews "ad-hoc risk assessment" by Google exists and is under review by the European Commission[54], but it has not been published, and its contents are not verifiable. Despite these signals, Gemini continues to occupy an ambiguous position. In Google's publicly available 2024 DSA systemic risk-assessment report, Google discusses generative AI extensively as a risk factor, but does not name Gemini, AI Overviews, or Search Generative Experience anywhere.[55]

The Gemini case thus illustrates a broader challenge where companies can strategically frame or delay integration of generative AI features to limit obligations. Google's earlier separation of Gemini from Google Search can be seen as a regulatory strategy, given that once generative outputs appear in the main interface, the platform's designation as a VLOSE becomes decisive.

ChatGPT: Standalone application

Despite ChatGPT's massive adoption, reaching around 10% of the world's adult population by mid-2025 and over 41 million monthly active users in the EU,[56] its regulatory classification remains unsettled[57]. Usage data shows that "seeking information" accounts for nearly 80% of interactions, positioning ChatGPT as a functional alternative to traditional search engines.[58] Yet, the DSA framework has not clearly determined whether ChatGPT qualifies as a hosting service, an intermediary, or a search engine. It has been reported earlier[59] that the Commission considers designating ChatGPT as a VLOP or VLOSE on the basis of its search functionality and its reach of over 45 million monthly active users.

However, when it comes to regulating standalone AI products under the DSA, definitional challenges persist. The DSA defines an intermediary service as an entity engaged in "mere conduit," "caching," or "hosting" services (DSA Art. 3(g)). Hosting services are defined as "the storage of information provided by, and at the request of, a recipient of the service"

54 Chee, Yun, and Jane Merriman. 2025. "Exclusive: Google's AI Overviews hit by EU antitrust complaint from independent publishers." *Reuters*, July 4, 2025. https://www.reuters.com/legal/litigation/googles-ai-overviews-hit-by-eu-antitrust-complaint-independent-publishers-2025-07-04/.

55 Google Ireland Limited. *Report of Systemic Risk Assessments 2024*. "DSA Risk Assessment 2024-08-28." January 2025. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2024-8-28_2024-8-28_en_v1.pdf

56 Alexander Bick, Adam Blandin, and David J. Deming, "The Rapid Adoption of Generative AI."

57 Politico. "The EU Can't Figure Out What to Do-about ChatGPT." *POLITICO Europe*,November 2025. https://www.politico.eu/article/eu-chatgpt-ai-digital-law-tech-openai-regulations-legal/

58 Alexander Bick, Adam Blandin, and David J. Deming, "The Rapid Adoption of Generative AI."

59 Luca Bertuzzi, "ChatGPT Faces Possible Designation," MLex, accessed September 20, 2025, https://www.mlex.com/mlex/articles/2332484/chatgpt-faces-possible-designation-as-a-systemic-platform-u

(DSA Art. 3(g(iii))). Whether ChatGPT fits this definition is contested.[60] Unlike a neutral host, ChatGPT generates outputs, but it also stores user inputs and uploaded files, suggesting a partial hosting function, including text and uploaded data (e.g., PDFs, images),[61] and it generates responses directly informed by this information, rather than operating independently of user contributions.[62]

Before the DSA harmonized the rules for intermediary services across the EU, the 2000 E-Commerce Directive applied, leading to regulatory fragmentation across member states and case law. To address such gaps, the DSA framework introduced a distinct definition of "online search engines" as a subset of "hosting services," although the former does not perfectly fit the definition of "hosting services."[63]

An online search engine in the DSA is defined as:

> "*an intermediary service that allows users to input queries in order to perform searches of, in principle, all websites, or all websites in a particular language, on the basis of a query on any subject in the form of a keyword, voice request, phrase or other input, and returns results in any format in which information related to the requested content can be found* (DSA Art. 3(j))."

The key ambiguity is whether the Commission would cover only ChatGPT's search feature under the DSA or consider the chatbot as a whole, inseparable from its search capability. The precedent of Zalando's challenge of the DSA on the basis of its hybrid business model suggests that such delineations can be tested in court.[64]

Despite limitations in formalistic approaches to intermediary services' liability laws in the EU, the current market trends clearly favor AI search as an alternative to and an evolution of traditional online search, with EU-level regulation so far failing to grasp all of its facets. Not only do web search features or the broader use of ChatGPT for search fit the definition of an online search engine, but there is a strong argument that it should be considered as one "when they are released in the market in a way that strongly resembles other intermediaries

60 Philipp Hacker, Andreas Engel, and Marco Mauer, "Regulating Chat-GPT and other Large Generative AI Models. In 2023 ACM Conference on Fairness, Accountability, and Transparency" (FAccT '23), June 12–15, 2023, ACM, New York, NY, USA. https://doi.org/10.1145/3593013.3594067

61 Paddy Leerssen, "Embedded GenAI on Social Media: Platform Law Meets AI law," DSA Observatory, accessed September 20, 2025, https://dsa-observatory.eu/2024/10/16/1864/

62 Mathias Vermeulen and Laureline Lemoine, "From ChatGPT to Google's Gemini," LSE Media Blog, accessed September 20, 2025, https://blogs.lse.ac.uk/medialse/2024/02/12/from-chatgpt-to-googles-gemini-when-would-generative-ai-products-fall-within-the-scope-of-the-digital-services-act

63 Beatriz Botero Arcila, "Is it a Platform? Is it a Search Engine? It's Chat GPT! The European Liability Regime for Large Language Models."

64 Cynthia Kroet, "Zalando's legal case to spark EU Commission online platform headache," Euronews, accessed September 20, 2025, https://www.euronews.com/next/2025/03/06/zalandos-legal-case-to-spark-eu-commission-online-platform-headache

covered by content moderation laws, such as search engines."[65] Considering ChatGPT's extensive user base, it may eventually be designated a VLOSE and subject to DSA obligations similar to those governing Copilot once it reaches the 45 million user threshold, which is likely already the case. Another important consideration would be whether the designation of ChatGPT Search as an online search engine under the DSA would grant it safe harbor, and how it would affect the provider's responsibility for the content it produces.

Lastly, the DSA arguably provides a relatively mature framework for addressing information-related harms. Extending selected provisions such as notice-and-action systems and trusted flaggers (DSA Art. 16) to ChatGPT[66] could offer a pragmatic way to mitigate misinformation and other systemic risks, even before a formal designation is finalized.

Beyond this interim logic, there is also a substantive benefit to applying the DSA alongside the AIA. The AIA is primarily concerned with the model layer, requiring providers to anticipate and mitigate risks at the level of system design. By contrast, the DSA applies to the *downstream stage of deployment*, where LLMs operate in real-world information environments. This dual coverage helps capture harms that emerge not only from how the model is trained but also from how it interacts with users, content ecosystems, and public discourse. In fact, as mentioned earlier, the GPT-5 System Card suggests that OpenAI already updates its models' safeguards and characteristics based on usage information; however, without regulatory requirements, this rests exclusively on the company's own discretion. For a product like ChatGPT, which is widely used for information-seeking and decision support, such downstream governance is essential to address risks to the information environment and democratic processes.

Regulation of ChatGPT under the AIA

ChatGPT is built on GPT-5, a model developed by OpenAI, and defined in the AIA as a General Purpose Artificial Intelligence (GPAI) model. As such, OpenAI must comply with obligations for GPAI models in the AIA. The AIA is a product liability law whose risk management approach incentivizes *ex ante* risk identification and mitigation. Article 51 of the AIA introduces obligations for the providers of "General Purpose AI Models with systemic risk", such as assessing and mitigating systemic risks at the union level. This requirement applies across the full life cycle of models with systemic risk (AIA Art. 55(b)). Counted among systemic risks in Recital 110 are "risks with reasonably foreseeable negative effects on public health, safety, democratic processes, public and economic security, fundamental

65 Beatriz Botero Arcila, "Is it a Platform? Is it a Search Engine? It's Chat GPT! The European Liability Regime for Large Language Models."

66 Philipp Hacker and Atoosa Kasirzadeh and Lilian Edwards, "AI, Digital Platforms, and the New Systemic Risk," https://ssrn.com/abstract=5475049

rights, and the society as a whole."[67] Moreover, systemic risks are acknowledged to potentially grow alongside the model's capabilities and reach, and to rise at any stage of its life cycle (AIA Art. 3(65)).

The AIA includes several indicators for classifying GPAI models with systemic risk (GPAISR). Among these indicators, "the cumulative amount of computation used for training measured in floating point operations (FLOPs)" is presented as the main indicator for this characterization (AIA Art. 51(2)). The choice of FLOPs as the main indicator, with its current threshold of 10(^25) as expressed in the law are criticized on multiple grounds. For instance, scholars such as Sara Hooker argue[68] that this indicator does not adequately correlate with the severity and probability of risk. Moreover, a dynamic interpretation of the AIA's focus on "most advanced" models might lead to widely-adopted, legacy models being omitted from the GPAISR category,[69] which might consequently lead to less safe but widely used models in the market due to lessening regulatory pressures.[70]

According to Epoch.ai's Tracking Large-Scale AI Models report,[71] OpenAI's GPT-4 exceeds the threshold of 10(^25) FLOPs, which means that it can be classified as a GPAI model with systemic risk. In case this indicator is contested or becomes less relevant by the time of application, the AIA also emphasizes "model reach" in Annex XIII, notably, whether a model has at least 10,000 registered business users or a large end-user base, as a factor in classification. Both aspects considered, there are reasonable grounds to anticipate GPT-5 being mandated to assess and mitigate systemic risks under the AIA.

Regardless, the Commission acts as the ultimate decision-maker on this classification. It retains the authority to amend and refine the criteria through delegated acts. Moreover, the Commission can decline to designate a model that ostensibly meets the criteria outlined in Annex XIII, based on a request from providers. Conversely, the Commission can designate a model following a qualified alert from the Scientific Panel, even if the model does not clearly meet the criteria (Art. 52).

The definitional scope and methodologies to assess and mitigate systemic risks are outlined under the General Purpose AI Code of Practice[72] until harmonized standards. Though

67 EU Artificial Intelligence Act, "Recital 110," Artificial Intelligence Act, accessed September 20, 2025, https://artificialintelligenceact.eu/recital/110/

68 Sara Hooker, "On the Limitations of Compute Thresholds as a Governance Strategy," ArXiV preprint, arXiv:2407.05694v2

69 Philipp Hacker and Matthias Holweg, "The Regulation of Fine-Tuning: Federated Compliance for Modified General-Purpose AI Models,"

70 Philipp Hacker and Atoosa Kasirzadeh and Lilian Edwards, "AI, Digital Platforms, and the New Systemic Risk."

71 "Tracking Large Scale AI Models," Epoch AI, accessed September 20, 2025, https://epoch.ai/blog/tracking-large-scale-ai-models#benchmarks-and-repositories

72 European Commission, "Contents Code GPAI," European Commission, accessed September 20, 2025,

https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

voluntary, the set of guidelines under the Code enables providers to demonstrate compliance with the Art. 53(4) and Art. 55(2) of the AIA starting from Aug 1, 2025. The Code focuses widely on AI safety-related risks and acknowledges broad risk categories such as "risks to public health, to safety, to public security, to fundamental rights, and to society as a whole. While GPAISR's "tendency to hallucinate, to produce misinformation, or to obscure sources of information" is mentioned in Appendix 1.3.2. Model Propensities section of the Code, misinformation itself is not outlined as a systemic risk category, neither in the Code nor the AIA. Therefore, this creates uncertainty whether the providers will assess information-related risks in their risk assessment process.[73]

Limits of product liability

In a product liability framework, manufacturers address safety requirements of their products by working with the fixed, tangible characteristics of the product before it reaches consumers.[74] The AIA's ex ante, product-safety logic is not fully adequate for LLMs integrated into multiple contexts and processes. Unlike fixed products, GPAI systems evolve through usage and API-based deployments. While the proactive risk management approach is useful, it is not, for instance, equipped to address information-related harms posed by LLM-based applications.[75] Similar to social media platforms, LLMs also pose risks related to the online information production and dissemination.[76]

The AIA framework primarily requires providers to address systemic risks in the pre-deployment phase. Although valuable, this approach may prove insufficient to risks that emerge from the deployment context. For example, ensuring that training data is representative is crucial to prevent bias and discrimination from permeating and amplifying across the AI value chain. However, GPAI systems are highly adaptable and can be deployed across diverse contexts. Even in the case of ChatGPT, where OpenAI serves as the model developer, host, and application developer, the current obligations would require the GPT-5 provider to evaluate the model against all potential high-risk scenarios outlined in Annex III, which is an expansive and potentially impractical mandate.

Post-deployment factors critically affect AI search system safety and reliability in ways that pre-deployment evaluation cannot fully anticipate. Research analyzing 1,178 safety papers from leading AI companies and universities reveals that corporate AI safety research

73 Philipp Hacker and Matthias Holweg, "The Regulation of Fine-Tuning: Federated Compliance for Modified General-Purpose AI Models."

74 Beatriz Botero Arcila, "Is it a Platform? Is it a Search Engine? It's Chat GPT! The European Liability Regime for Large Language Models."

75Ibid.

76Ibid.

disproportionately focuses on pre-deployment risks while overlooking real-world harms, including information-related risks such as misinformation.[77]

Critical post-deployment components affecting system behavior include orchestration primitives, routing information among users and external systems, data-retrieval layers like retrieval-augmented generation (RAG) supplying knowledge beyond training data, safety services enforcing company policies through moderation filters, and evaluation systems tracking quality and user feedback. These components interact in complex ways that pre-deployment testing cannot fully capture.

Furthermore, voluntary safety commitments do not present as strong a mechanism for compliance: UK parliamentarians accused Google of releasing Gemini 2.5 Pro without adequate safety testing documentation,[78] while ongoing discussions consider the company's delaying of AIA implementation timelines and regulatory simplification. Corporate incentive structures favor rapid deployment over comprehensive downstream risk monitoring, creating systematic gaps in post-market safety oversight.

The distinction between "moderating behavior" (AIA focus) and "moderating content" (DSA approach) suggests the value of viewing these frameworks as complementary rather than competing for regulating AI search systems. The AIA's anticipatory governance through pre-deployment evaluation should be paired with DSA-style operational oversight addressing information-related systemic risks as they emerge.

ChatGPT's widespread use as a search engine, despite remaining a "standalone product," creates regulatory gaps that selective DSA application could address[79]. Legal scholars propose applying DSA content moderation tools such as notice-and-action mechanisms and trusted flagger systems to ChatGPT, adapting platform governance concepts to AI-generated content.[80] This hybrid approach would leverage the DSA's extensive experience with information-related systemic harms while maintaining the AIA's systematic risk assessment framework.

Current transparency mechanisms prove insufficient for AI search systems operating at ChatGPT's scale and impact level. While OpenAI publishes internal audit reports on malicious use monitoring and information ecosystem impacts, independent verification remains impossible due to a lack of third-party access to usage data and methodologies. The AIA's

77 Strauss, Ilan, et al. "Real-World Gaps in AI Governance Research." arXiv, May 2025, arXiv:2505.00174v2, https://arxiv.org/html/2505.00174v2

78 Harry Booth, "Exclusive: 60 U.K. Lawmakers Accuse Google of Breaking AI Safety Pledge," The Times, accessed September 20, 2025, https://time.com/7313320/google-deepmind-gemini-ai-safety-pledge/

79 Philipp Hacker, Andreas Engel, and Marco Mauer, "Regulating Chat-GPT and other Large Generative AI Models."

80 Ibid.

information requests and systemic risk investigations operate primarily at Commission and Scientific Panel discretion, limiting accessibility for public interest research. This contrasts with DSA requirements for VLOPs and VLOSEs to publish systemic risk assessments and provide researchers with data access, enabling independent verification of risk mitigation effectiveness.

Systematic researcher access to AI search is crucial for understanding information-related systemic risks in contexts such as public health and democratic processes. Recent calls for subsidized API access and usage data availability reflect growing recognition that closed foundation models require external scrutiny to ensure adequate risk assessment and mitigation.[81]

81 Gabriel Nicholas, "Grounding AI Policy: Towards Researcher Access to AI Usage Data," The Center for Democracy & Technology, accessed 20 September, 2025, https://cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/; Esme Harrington and Mathias Vermeulen, "External researcher access to closed foundation models: State of the field and options for improvement,"The Mozilla Foundation, (2024), https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/External-researcher-access-to-closed-foundation-models.pdf

Toward Integrated AI Search Governance

The current European regulatory landscape creates gaps when applied to AI search systems. The DSA's focus on ongoing operational oversight of user-generated content and the AIA's emphasis on pre-deployment product safety create a regulatory divide that AI search systems traverse uneasily. Our case study analysis demonstrates how different LLM-deployment configurations: embedded, semi-integrated, or standalone, face varying regulatory treatment despite their similar functional roles as information intermediaries.

The regulatory treatment of Microsoft Copilot, Google Gemini, and OpenAI's ChatGPT illustrates discrepancies between existing regulatory frameworks. Copilot's clear integration into Bing Search places it squarely under DSA oversight. Gemini's ambiguous positioning demonstrates how companies can strategically structure deployments to limit regulatory exposure. ChatGPT's standalone status, despite widespread use for search purposes, reveals gaps in current definitions of online intermediary services.

Regulatory coordination is needed to address AI search systems comprehensively rather than allowing governance gaps to emerge across the AI value chain and through technical architecture choices. The complementary nature of anticipatory and reactive moderation approaches aligns well with the AIA's ex-ante product safety focus and the DSA's ex-post operational oversight, suggesting opportunities for integrated governance rather than competing regulatory regimes.

The concentration of AI search capabilities among a small number of global technology companies raises concerns about information diversity and market competition similar to those that have long surrounded traditional search engines. However, AI search systems' ability to synthesize and generate responses rather than simply rank existing sources creates new forms of informational power that may prove as consequential as traditional search dominance.

The opacity of AI search systems' decision-making processes, from training data composition to real-time content generation, undermines traditional accountability mechanisms that rely on source transparency and editorial oversight. Without adequate governance frameworks, these systems risk becoming black boxes that shape public knowledge without sufficient democratic input or oversight.

Policy recommendations: a path forward

To govern AI search systems effectively, we need several coordinated policy interventions that build on existing regulatory regimes while addressing their unique characteristics. For this, we recommend the following policy considerations:

European regulators should clarify the existing DSA obligations that ca be applied AI search (e.g., notice–and–action mechanisms, trusted flaggers, systemic risk assessments), and how it affects the "intermediary liability" provision under the DSA.

The European Commission should develop guidance on how the AIA and DSA will apply to AI search, given that effective governance requires both pre–deployment behavioral alignment and ongoing operational oversight. This coordination should address responsibility allocation between model developers and AI system deployers and operators.

AI search systems should face mandatory disclosure requirements which offer clarity on training data composition, behavioral alignment processes, source attribution methods, and content generation mechanisms. Public oversight can be facilitated through these requirements while protecting legitimate business interests and technical security concerns.

Independent researchers should receive structured access to usage data, behavioral patterns, and the effectiveness of risk mitigation from AI search systems through mechanisms similar to those provided in the DSA. This is key to public–interest investigations into information–related systemic risks and their democratic impacts.

Because AI search systems have global reach and impact on information ecosystems, European regulatory frameworks should coordinate with international partners on technical standards, enforcement cooperation, and research collaboration while maintaining regulatory autonomy over European information environments.

As AI search systems continue evolving, they require regulatory frameworks that can adapt to technological changes while maintaining consistent principles for information governance. The anticipatory and reactive moderation framework, proposed in this report, provides conceptual stability while allowing for technical adaptation as AI systems become more sophisticated.

Future research should examine the long–term impacts of AI search systems on information diversity, source credibility, and democratic discourse. Longitudinal studies of user behavior, information quality, and market dynamics will be essential for understanding whether

current governance approaches adequately protect public interests while enabling beneficial innovation.

Appendix

Terminology

The terminology used in this report is inconsistent and evolving. It is hard to find authoritative definitions that satisfy technical, academic, and policy audiences at once. While some terms are used interchangeably, we attempt to draw loose boundaries while referring to a set of underlying technologies, AI systems, products, and features within the broader area of "AI search." Therefore, this section should be understood as a reading guide rather than an authoritative definition section.

Search engine – A software system that provides information, usually in the form of web links, in response to users' queries. It is based on crawling, indexing, and ranking web-based sources and retrieving relevant web pages when queried by the user.

LLM – A large language model (LLM) is a transformer-based model trained on massive amounts of text data that generates text in response to a broad range of inputs (also referred to as "prompts"). LLMs are not explicitly programmed for a specific task and generate text by predicting the next token. In this report, we often refer to "large" language models per their number of parameters that are commercialized (e.g., Gemini, OpenAI's GPT-5, Anthropic's Claude). For instance, GPT-5 is an LLM and the model that underpins the popular conversational chatbot ChatGPT.

General-purpose AI model / Foundation model / Generative AI – Also referred to as a foundation model (FM), a general-purpose AI model (GPAI) is trained on large amounts of data that can be adapted to a broad range of tasks and applications, such as text, image, and audio generation. A GPAI model includes LLMs, but is not limited to them. This term is more prominent in policy terminology and is notably defined in the AI Act.

Generative AI – A class of artificial intelligence systems that generate content based on patterns acquired from the training stage on large datasets The content is usually outputted as a function of the provided input (such as a 'prompt' that the user is in control of) and a multitude of parameters (most of which usually remain outside of the user's direct control, such as a 'temperature' or the 'randomness' of the output).

AI chatbot – A user-oriented version of an LLM model; the user does not have full access to the LLM, only to a fraction of its deployment within the chatbot. Some chatbot platforms might choose to hide parts of the model's output (e.g., reasoning outputs, dangerous content). The chatbot may be capable of engaging in search (generally using a search engine), processing the results in a way that can vary from listing and sorting them to summarization, research (recursive search & processing of the results), and manipulation of the results, based also (but not exclusively) on the user's request.

Model vs AI system – An AI model is defined as a computational representation that encompasses processes, objects, ideas, people, and interactions. Foundation models (FMs) and large language models (LLMs) are examples of AI models, but an AI system is the complete deployment or product into which the model is integrated, along with the necessary logic and infrastructure to operate and interact with users and the environment.

AI value chain – The concept of the AI value chain describes the stages and actors involved in the development, provision, and use of AI components and systems, often emphasizing the relationship between foundation models and their ultimate deployment in specific applications.

Upstream vs downstream application – In the context of AI, the upstream usually refers to the initial development and supply of the foundational component, i.e., the general-purpose AI model. Downstream systems/applications are those that integrate a general-purpose AI model. For instance, we can refer to a chatbot, which is a conversational AI system, as a downstream application of an LLM.

Generative AI features – The integration of generative AI into an existing product involves incorporating a trained AI model (often a foundation model or large language model) to extend the product's capabilities, typically enabling the creation or augmentation of content or providing advanced, natural language-driven functionality. In this report, we classify a "generative AI feature" as such when it is visibly embedded in the user interface of an existing product, as an additional or separate functionality that users can use to enable new functions or improve existing ones. Examples include generative AI features integrated into non-AI products, such as Google AI Overviews, Adobe Photoshop, and integrated image generation models to allow for image generation or editing.

LLM-powered search – A search engine that makes use of generative AI features to augment the search results. It is supposed to assist the user in answering a query either by answering the query directly, summarizing the results, or a mixture of the two. In this approach, the foundation model is directly integrated into the software package or backend of an existing consumer product or via an API, often without requiring the user to have a technical understanding of the AI operating the feature.

Chat-based search – A search performed in engagement with an AI chatbot interface. The chatbot autonomously interacts with a search engine to collect results, processes them, and presents the user with an output that is intended to integrate a portion of the search results' content. The user is not in control of the search and the filtering of the results.

AI search – A digital search paradigm, where search outputs are provided by or supplemented through engagement with a chat-based search interface and/or generative AI features that include LLM-powered search and chat-based search described above.

Traditional search – Search outputs provided by a search engine that only uses algorithmic or "computational," reproducible, repeatable, rule-based techniques to find and sort results

(web pages) matching the user query.

Leading AI chatbots: usage and functionalities

Tool	Average daily visits	LLM family	Company	Search engine integration
Microsoft Copilot	2.4M–345M average daily visits[82]	GPT-4	Microsoft & OpenAI	**Yes,** integrated with Bing
ChatGPT	177.42M average daily visits[83]	GPT-4o GPT 5	OpenAI	**Standalone** + available as a separate search model that relies on the Bing Search index
Meta AI	>40M average daily visits[84]	Llama	Meta	**Standalone** + integrated across Meta apps (Facebook, Instagram, WhatsApp)
Grok	16.5M average daily visits[85]	Grok?	xAI	**Standalone** + direct interface to X's social media content
Deepseek	16.5M average daily visits[86]	DeepSeek-V DeepSeek R	Deepseek	**Standalone**
Gemini	10.9M average daily visits[87]	Gemini	Google	**Yes,** embedded into AI Overviews of Google Search Engine + the "AI Mode"

82 Kyle Wiggers, "ChatGPT Isn't the Only Chatbot that's Gaining Users," TechCrunch, accessed May 20, 2025, https://techcrunch.com/2025/04/01/chatgpt-isnt-the-only-chatbot-thats-gaining-users/; Fabio Duarte, "Number of ChatGPT Users (October 2025)" Exploding Topics, accessed May 20, 2025, https://explodingtopics.com/blog/chatgpt-users; Microsoft hasn't publicly stated how many total users Copilot has given that it is now integrated within Microsoft 365 suite of a user base that is estimated at nearly 345 million

83 Arooj Ahmed, "ChatGPT Usage Statistics: Numbers Behind Its Worldwide Growth and Reach (September, 2025)" Digital Information World, accessed May 20, 2025, https://www.digitalinformationworld.com/2025/05/chatgpt-stats-in-numbers-growth-usage-and-global-impact.html

84 The 40 million figure comes from an estimation given that Meta AI has been integrated with other Meta apps and is accessed mostly through them. The only official statement on the approximate number of users was given at Q4 2024 earnings call, announcing more than 700 million monthly active of Meta AI

85 Kyle Wiggers, "ChatGPT Isn't the Only Chatbot that's Gaining Users."

86 Kyle Wiggers, "ChatGPT Isn't the Only Chatbot that's Gaining Us.ers."

87 Kyle Wiggers, "ChatGPT Isn't the Only Chatbot that's Gaining Users." However, the company hasn't publicly stated how many people actually make use of Gemini within the company's portfolio of products.

Claude	3.3M average daily visits[88]	Claude 3.5 Haiku Claude Sonnet Claude Opus	Anthropic	**Standalone**
Perplexity	2–3M average daily visits[89]	GPT-4.1 o4-mini, Claude 4.0 Grok 3 Beta Gemini Sonar (based on Llama 3.3) R1 1776 (based on DeepSeek R1)	Perplexity AI	**Standalone**
Mistral	No data (we can estimate approx. 333,000 daily visits[90])	Mistral	Mistral AI	**Standalone**

Table 1. Table of the most popular LLM chatbots based on data from May 2025

How AI search differs from traditional search

haracteristic of the technical system	Traditional search	AI search	Amplified risks	New risks
Web-based sources	Yes	Yes	Biases	n/a
Retrieval-ugmented generation (RAG)	No	Yes	Disinformation, harmful and illegal content	n/a
Training data	No	Yes	n/a	Outputs based on outdated, biased, and limited data
				Factual errors ("hallucinations")

88 Kyle Wiggers, "ChatGPT Isn't the Only Chatbot that's Gaining Users."

89 Eamonn O'Raghallaigh, "The Key SEO Trend in 2025 – The Rise of Conversational AI Search," Digital Strategy, accessed September 20, 2025, https://digitalstrategy.ie/insights/the-key-seo-trend-in-2025-the-rise-of-conversational-ai-search/

90 Mistral AI, "Tech Details, Crunchbase, accessed September 20, 2025, https://www.crunchbase.com/organization/mistral-ai/technology

Ranking sources	Yes	Yes	SEO gaming into data poisoning	n/a
Assigning relevancy scores to sources	Yes	Yes	Further simplification and limitation of the knowledge pool	Ultimate single answer
Prediction-based generation	No	Yes	n/a	Factual errors ("hallucinations")
Personalization	Yes	Yes	Biased results delivered based on a broader users' data profiling	User-specific output personalization, tailoring to users' (implicit) biases, and chat personalization
Source attribution	Yes	Partially	n/a	No clear source attribution, lack of transparency
Organic search results intertwined with sponsored content and ads	Yes	Yes	Increased collapse in transparency between organic and sponsored results content in the search outputs	n/a
Single output in a query-receipt exchange	No	Yes	n/a	Plausible-looking factual errors an authoritative "stance"

Search engines & existing issues

The challenges of algorithmic sorting, ordering, and ranking of search results

Search engines, such as Google and Bing,[91] **are information retrieval systems that rely on algorithmic indexing and ranking processes to locate, rank, and present web-based sources in response to users' queries.** Search engines, driven by their broad definition of "relevance and often an advertisement-based business model, control information flows and content discoverability. They blend knowledge access with sponsored content and personalized advertisements, solidifying their dominance in the information economy. They surface certain content to users, while drowning others.[92] The impact can make a company succeed or die (e.g., Mappy). This effect has led to the emergence of an entire industry of Search Engine Optimization (SEO), estimated at USD 74 billion in 2024.[93]

Depending on the context, some sources might be moderated to always be considered particularly authoritative, such as governmental and World Health Organization websites during the COVID-19 pandemic. Additionally, search results include non-organic results, such as sponsored content, advertisements, and, as is the case with Google, the company's own properties (e.g., links to YouTube and Google Maps). Search engines, such as Google, can also push sponsored content among organic results and promote their own products without disclosing it in an outward manner.[94] For example, to further consolidate its dominance, Google has introduced search-related features such as "direct answers" and "people also ask" that are occupied by Google properties at 42%, often "pushing down the organic results.[95]

Efficient information seeking online necessitates a deployment of some sorting and ordering mechanisms. In search engines, users accept a partial delegation of knowledge selection and ordering to algorithmic processes, where the judgment of relevance is predetermined

91 In the following paragraphs, we refer mostly to practices used by Google as it has been the most popular search engine in the world with a monopoly influence to shape online search.

92 Pascal Jürgens and Brigit Stark, "The Power of Default on Reddit: A General Model to Measure the Influence of Information Intermediaries," Policy & Internet, 9: 395–419, 2017, DOI:10.1002/poi3.166; Michael Latzer, Katharina Hollnbuchner, Natascha Just, Florian Saurwein, "The economics of algorithmic selection on the Internet," In: Johannes M. Bauer and Michael Latzer (ed.), Handbook on the Economics of the Internet p. 395–425 (Edward Elgar Publishing, 2016).

93 Max Roslyakov, "SEO Market Stats (2024)," Xamsor, accessed 20 September 2025, https://xamsor.com/blog/seo-market-stats/

94 Michael Luca, Tim Wu, Sebastian Couvidat and Daniel Frank, "Does Google Content Degrade Google Search?" Experimental Evidence, Harvard Business School NOM Unit Working Paper No. 16-035, 2015, https://scholarship.law.columbia.edu/faculty_scholarship/1931; Daniel A Crane, "After Search Neutrality: Drawing a Line Between Promotion and Demotion," I/S: J. L. & Pol'y for Info. Soc'y 9, no. 3 (2014): 397-406

95 Adrianne Jeffries and Leon Yin, "Google's Top Search Result? Surprise! It's Google," The Markup, accessed May 20, 2025, https://themarkup.org/google-the-giant/2020/07/28/google-search-results-prioritize-google-products-over-competitos

by the search engine. Search engines' "particular *knowledge logic*"[96] can favor and impact what users encounter (and why) when searching for information. The order of results, as well as functionalities such as Google's autocomplete, offer ease and simplicity in exchange for certain delegation of (web) knowledge curation.

Moderation of search results in search engines

What Google shows (the order of results and the results themselves) is an outcome of algorithmic selection, human refinements, and moderation processes. Google combines the notion of relevance with some degree of personalization, accounting for a user's datafied profile and physical location, search history, and device specificities, with previous manual adjustments of results by Google's human moderators, and A/B testing. On the whole, this approach combines a range of algorithmic outputs with some human choices and moderation. The consequent search output may result in biased results,[97] often reflecting the user's own implicit biases embedded in the choice of search queries.[98]

Google has participated in the perpetuation of inadvertent biases in both search results and its autocomplete suggestions; for instance, searching for names commonly given to African Americans produced significantly more advertisements for criminal record searches, while terms related to Asian and Black girls mostly led to porn.[99] It has been proven that biases in search suggestions and search results influence people's views on vaccinations,[100] as well as the voting preferences of undecided voters after just one search[101] (a phenomenon termed "search suggestion effect"[102]). Google improved its search moderation and autocomplete suggestions after external criticism, though English queries see stricter, broader moderation than other languages.[103]

96 Tarleton Gillespie, "The Relevance of Algorithms," in Tarleton Gillespie, Pablo J. Boczkowski, Kirsten A. Foot (eds) Media Technologies: Essays on Communication, Materiality, and Society (Cambridge, MA: The MIT Press, 2014), 168

97 Eli Pariser, The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think (New York: Penguin, 2011); Mario Haim, Andreas Graefe, and Hans-Bernd Brosius, "Burst of the Filter Bubble? Effects of personalization on the diversity of Google News," Digital Journalism, 6 (3), 2017: 330–343

98 Axel G. Ekström, Guy Madison, Erik J. Olsson, and Melina Tsapos, "The search query filter bubble: effect of user ideology on political leaning of search results through query selection," Information, Communication & Society, 27 (5), 2023: 878–894

99 Safiya Umoja Noble, Algorithms of Oppression: How Search Engines Reinforce Racism (New York: NYU Press, 2018); Astrid Mager, Ov Cristian Norocel, and Richard Rogers, "Advancing search engine studies: The evolution of Google critique and intervention," Big Data & Society, 10 (2), 2023

100 Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto, "The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating Google output," Journal of Medical Internet Research, 16 (4), 2014: e100

101 Robert Epstein and Ronald E. Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections," Proceedings of the National Academy of Sciences, 112 (33), 2015: E4512–E4521

102 Robert Epstein, Savannah Aries, Kally Grebbien, Alyssa M. Salcedo, and Vanessa R. Zankich, "The Search Suggestion Effect (SSE): How Autocomplete Search Suggestions Can Be Used to Impact Opinions and Votes," Computers in Human Behavior, 160C, 2024

103 Richard Rogers, "Algorithmic probing: Prompting offensive Google results and their moderation," Big Data & Society,10(1), 2023

Search engines have been mostly criticized for attributing credibility and visibility to problematic or harmful sources and suggestions that have already been published and shared on the web by users. AI search expands the margin of risk by generating outputs anew using users' queries, the training corpus, and web-based content as contextual pointers, as well as moderation loops set by the companies behind AI models and chatbots (see "Rethinking Moderation in AI Search" section).

Reintroducing AI chatbots from the search angle

LLMs' training datasets as preselected knowledge representations

The vast training datasets, predominantly scraped from the internet and reflecting Western, English-language, and dominant cultural perspectives, introduce biases into LLMs' search-generation outputs.[104] The training process for an LLM maps the relationships between words and concepts based on their co-occurrence patterns in the training set. This training set is the primary source of information, and the more an idea is represented within it, the more likely it is to appear in the model's future outputs. The training process for an LLM has several stages that contribute to the available knowledge pool and can be further enriched by retraining the model on updated or specialized datasets.

A single model can generate deterministic, reproducible results, yet the flexibility and indeterminism of its outputs tend to be more desired by both providers and users. For example, the same prompt might yield a different output across time, partially as a result of the company behind it tempering with the chatbot's preferred outputs (e.g., by adjusting the temperature, meaning the level of randomness in the response), as well as new moderation adjustments, and user-chatbot personalization. In the context of search engines, certain countermeasures have been developed over their 20 years of deployment that make it possible to remove (or block) websites that are deemed problematic or harmful. However, so far, it appears that this cannot be done as efficiently and definitely in LLMs given their training datasets.[105] This technical difference proves crucial when we consider the applicability of moderation to LLM-generated information that seems plausible but is misleading or inaccurate.

Model designers can influence the model's "worldview" by choosing which data sources to include and how to weight them, for instance, by adding Reddit for folk theories or Wikipedia for encyclopedic knowledge. Each such data source can be given more or less

104 Cameron Pattison, Vance Ricks, and John Wihbey, "How AI-Driven Search May Reshape Democracy, Economics, and Human Agency," Tech Policy Press, accessed September 20, 2025, https://www.techpolicy.press/how-ai-driven-search-may-reshape-democracy-economics-and-human-agency/.

105 Emma R. Murphy, Nadia Madkour, Deepika Raman, Kelsey Jackson, and Jesse Newman, "Survey of Search Engine Safeguards and their Applicability for AI," 21

weight in the training set. By choosing to include and weigh more or less certain types of documents, with certain topics, viewpoints, or cultural representations, model designers can influence the model's representation of the world and its future response. This can be seen as another similarity between AI search and search engines: just as outputs in LLMs can be influenced by model designers, the ranking mechanisms in search engines also employ machine learning systems, where developers can intervene to adjust weights as part of a moderation task.

Generation and hallucination versus retrieval of information

While search engines index and retrieve content from the web, LLMs do not directly retrieve content but rather generate it from learned patterns and contextual input.[106] Even when LLM-powered chatbots have integrated retrieval-augmented generation (RAG) in an attempt to improve outputs, the information LLMs generate is freshly created, as "current LLMs' responses are created by 'forecasting token sequences with the highest likelihoods.'"[107] In other words, even if the RAG provides context for the AI chatbot, the model still has to process it to generate stochastic output (rather than retrieving, scraping, or copying relevant information). It must be noted that an LLM-based application could be instructed to report contextual content word-for-word (e.g., by utilizing the additional tool to perform a copy-and-paste operation). Yet, in principle, LLM-based services do not retrieve outputs; they generate them.

LLMs' outputs are mostly accurate, which makes the occasional but inevitable factual errors plausible-looking and more dangerous. Widely referred to as "hallucinations", the term designates "content that diverges from the user input, contradicts previously generated context, or misaligns with established world knowledge."[108] [109] Hallucinations are ultimately factual errors which are "a fundamental feature of LLMs beyond training data" that cannot be fully eradicated,[110] [111] an unavoidable feature of even the state-of-the-art AI chatbots that even companies such as OpenAI begin to openly recognize.[112] LLMs might also "snowball" from prior errors, which means "over-commit[ting] for self-consistency rather than recovering from errors" even when an error in the generated text strings is detected.[113]

106 Emma R. Murphy, Nadia Madkour, Deepika Raman, Kelsey Jackson, and Jesse Newman, "Survey of Search Engine Safeguards and their Applicability for AI," 21

107 Haoyi Xiong et al., "When Search Engine Services meet Large Language Models: Visions and Challenges."

108 Yue Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," ArXiv, arxiv.org/pdf/2309.01219, 2024, 1

109 Yue Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," 4

110 Ibird.

111 Jiawei Yao, Kai Ning, Zhiyuan Liu, Maosong Ning, and Ling Yuan, "LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples," ArXiv, 2024, arXiv:2310.01469

112 OpenAI , "Why Language Models Hallucinate," OpenAI, accessed September 20, 2025, https://openai.com/index/why-language-models-hallucinate/

113 Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models,"10

Furthermore, LLMs might output errors (outdated or incorrect information) as well as generate new errors and biases (to "make up information," see e.g., Google's "AI Overviews" making up definitions for proverbs as reported by Wired).[114] Factual errors can sometimes be attributed to sources whose content has been misquoted by AI search, causing reputational risk or extending plausibility to incorrect statements.[115]

Inconsistent and disappearing source attribution

Unlike traditional search results with clear source attribution, AI chatbots often synthesize multiple sources in outputs without transparent source attribution. Users receive authoritative-sounding answers without understanding which sources informed specific claims or how conflicting information was resolved, thereby increasing the already widespread risk posed by clickbait and misleading websites that aim to game SEO rankings. The scale of the problem creates a new "attribution gap:" in most cases, chatbots provide little or no clickable source attributions or cite less than 50% of the web pages they visit to scrape relevant information.[116] There is little to no explainability and transparency (let alone outside scrutiny) to assess why, e.g., within an AI Overview, certain sources and/or facts are excluded or included, and where they were sourced from,[117] breaking away from any right to explanation. The deepening obstacles in source attribution and credibility posed by AI search might undermine users' ability to assert and evaluate the search results, a problem that further underscores the pressing need for media literacy.

The inevitable future of sponsored content and ads

In AI search outputs, the lack of clarity about source attribution calls into question the level of transparency in disclosing sponsored content and targeted advertisements. Given that the business model of most online search infrastructure is advertising, the introduction of advertisements and sponsored content into AI Overviews and LLM-based outputs is rather unavoidable. Once AI summaries and AI chatbots begin to output answers containing sponsored content and paid-for ads, the impact of AI search on transparency and organic information ordering and accessibility will be more concerning; Google has already started selling ads against its AI Overviews.[118]

---

114 Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜" in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (New York: Association for Computing Machinery, 2021), 610–623; Brian Barrett, "Google AI Overviews Meaning," Wired, accessed September 20, 2025, https://www.wired.com/story/google-ai-overviews-meaning/?sp=e823d799-2747-40c0-ac95-5a5f552fc28f.1747840680404

115 Salvatore Romano, et al., "Prompting Elections: The Reliability of Generative AI in the 2023 Swiss and German Elections," AI Forensics & Algorithm Watch (2023), https://aiforensics.org/work/bine-chat-elections

116 Ilan Strauss, Jangho Yang, Tim O'Reilly, Sruly Rosenblat, and Isobel Moure, "The Attribution Crisis in LLM Search Results."

117 Kathleen A. Creel, "Transparency in Complex Computational Systems," Philosophy of Science, 87 (4), 2020: 568–589

118 Cameron Pattison, Vance Ricks, and John Wihbey, "How AI-Driven Search May Reshape Democracy, Economics, and Human Agency."

Answering with more authority and less accountability

This single, direct answer approach creates a problematic simplification and limits information accessibility and the knowledge pool. The "direct answer" strategy that Google has been a leader in pursuing is reflected in how most AI chatbots answer queries. However, in AI search, "a frictionless sequence of query, receipt, and tacit acceptance"[119] bears no alternatives: no other links and opinions that the algorithmic processes excluded from the summary. This is more alarming given the authoritative "stance" most such answers take; chatbots' "outputs tend to drop qualifiers like 'might' and 'may,' and thus risk dampening users' critical reflexes even further, producing an epistemic vulnerability in which individuals encounter content whose lineage is opaque yet feel diminished incentive to interrogate its accuracy."[120] Users might fall into the fallacy of believing the single answer outputted by a chatbot or an AI summary as an ultimately correct and exhaustive one.

The algorithmic authority to select, process, and generate question–answer deliverables introduces yet another and more severe algorithmic layer of knowledge delegation and ordering. This layer introduces a further shift away from accountability and authority; AI search introduces both a level of "model opacity and institutional opacity"[121] which was first presented with search engines. Users, civil society, and governance parties are also unable to verify or exert agency over how and why knowledge is represented by AI search, as they have no access to the reasoning behind or the content of all outputs generated by a model or chatbot embedded in/as the search functionality.

Unmatched level of personalization

Finally, the level of algorithmic authority that AI search is prescribed is also alarming given its unmatched output personalization, tailoring to users' (implicit) biases, and partaking in delusional information spirals. The companies behind AI chatbots, such as Anthropic and OpenAI, but also Google in its approach to Gemini, are openly stating that ultimate personalization is their goal for further product development.[122] The delegation of knowledge to (personalized) algorithmic processes without equitable transparency and accountability measures has far-reaching cognitive and epistemological impacts. The threat of delusional spiraling, such as presenting information that enforces the user's (implicit)

119 Cameron Pattison, Vance Ricks, and John Wihbey, "How AI-Driven Search May Reshape Democracy, Economics, and Human Agency."

120 Ibid..

121Ibid.

122 Mark MacCarthy, "The Privacy Challenges of Emerging Personalized AI Services," Tech Policy Press, accessed September 20, 2025, https://www.techpolicy.press/the-privacy-challenges-of-emerging-personalized-ai-services/

biases, assumptions, fears, and hopes, is especially harmful, and even life-threatening to children, teenagers, and people with mental health predispositions.[123]

123 Kashmir Hill and Dylan Freedman, "Chatbots Can Go Into a Delusional Spiral. Here's How It Happens," The New York Times, accessed September 20, 2025, https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html

Moderation instances in AI search

Model training as editorial foundation

The pre-training of a "raw" LLM is a long and expensive process, typically performed only once. Through this process, the model learns the statistical distributions of words in a vast training text corpus. The curation of this training dataset represents the foundational governance layer, determining the foundational knowledge on which all subsequent reasoning will be based.

The training dataset serves as the primary information source from which a knowledge map is derived. The more an idea appears in the training set, the more likely it will appear in future outputs. Companies can weigh different data sources, effectively making editorial choices about which topics, viewpoints, or cultural representations receive emphasis.

This makes the training dataset composition particularly sensitive. The weighting of different elements is among the most tightly guarded secrets of AI companies, even those releasing open-source models, because these choices fundamentally shape the model's representation of the world and future responses.

Fine-tuning and RLHF

Fine-tuning and reinforcement learning from human feedback (RLHF) further shapes model behavior after initial training. Fine-tuning uses smaller, curated datasets to adjust model responses, while RLHF incorporates human preferences by having annotators label preferred outputs for the same prompts. Because fine-tuning of data isn't diluted through the use of massive training datasets, these interventions have stronger effects on model behavior (though these effects might be superficial, as what is considered bias might be better concealed in the models' outputs).

Red teaming

Red teaming provides an adversarial engagement and evaluation of model risks through controlled examination by expert groups. This process stress-tests models using prompts "that carry potential societal, economic, infrastructural, or security risks, including those involving discrimination, sexual violence, physical violence, self-harm, and other categories drawn from harm-related taxonomies."[124] As such, red teaming involves creative

124 Bernhard Rieder and Yarden Skop, "The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API," Big Data & Society, 8(2), 2021: 20539517211046181

engagement with black boxes," experimenting with both intended and unintended use cases that can lead to a reiteration of the system in order to improve it.

System prompts as real-time control

System prompts instruct the model on how to behave in response to the queries, sometimes overwriting the prompts, e.g., by adding behavioral instructions before they reach the AI model. When users submit queries, they aren't transferred directly to the language model; instead, they're expanded with additional instructions, such as "answer the following question with a polite and formal tone" or specific content boundaries.

Unlike training and fine-tuning modifications, system prompt updates can be deployed rapidly, making them the preferred mechanism for addressing emerging behavioral issues or making quick fixes to chatbot responses. The Grok political bias shift exemplifies this layer's power and opacity. System prompts are part of every conversation, and, as they're hidden from the user, there is a risk that they could be personalized

Retrieval-augmented generation as information gatekeeping

For all web-connected AI search systems, traditional search engines are used to some extent in order to retrieve relevant, new information. Retrieval-augmented generation (RAG) can be further enriched by user queries with additional information and documents to provide context for response generation. The RAG process involves searching for relevant snippets in a document database, making it similar to search engine functionality.

The content selection process through RAG highly influences chatbot responses, making the design of this selection process a form of moderation operating through three steps:

Indexing determines which documents are available for retrieval. If a document or internet page isn't indexed by the internal system, it won't be considered in context enrichment.

Ranking assigns relevance scores to indexed documents related to user queries. Similar to traditional search engines, RAG systems rank candidates based on relevance of the query in relation to the document's body, mirroring to a certain degree PageRank-style approaches.

Context Selection chooses specific text snippets from ranked results for incorporation into responses. This process determines how faithfully ideas are reflected in answers and whether links to sources are included, going beyond traditional search engine presentation of ranked results.

Input/output filtering as safety mechanisms

Input and output filters provide final safeguarding layers to prevent conversations on sensitive, dangerous, or illegal topics. Input filters evaluate user queries for problematic topics or keywords, deflecting queries that exceed risk thresholds with generic responses like "Sorry, I cannot respond to this" or redirecting users to appropriate resources.

Output filters evaluate AI responses before delivery to users, checking for offensive, illegal, or dangerous content. Despite all upstream moderation steps, model designers cannot guarantee appropriate responses to all prompts, making output evaluation essential for identifying and potentially redacting problematic content.

The election content blocking cases demonstrate both input filtering (keyword detection in queries) and output substitution (replacing generated responses with standardized disclaimers).

Bibliography

Aaron Chatterji et al., "How People Use ChatGPT. Working Paper 34255" National Bureau of Economic Research, accessed September 20, 2025, http://www.nber.org/papers/w34255

Adrianne Jeffries and Leon Yin, "Google's Top Search Result? Surprise! It's Google," The Markup, accessed May 20, 2025, https://themarkup.org/google-the-giant/2020/07/28/google-search-results-prioritize-google-products-over-competitos

Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto, "The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating Google output," Journal of Medical Internet Research, 16 (4), 2014.

Alex Heath, "OpenAI wants ChatGPT to be a 'super assistant' for every part of your life," The Verge, accessed 20 May, 2025, https://www.theverge.com/command-line-newsletter/677705/openai-chatgpt-super-assistant

Alexander Bick, Adam Blandin, and David J. Deming, "The Rapid Adoption of Generative AI," National Bureau of Economic Research, Technical Report 32966, 2024, https://www.nber.org/system/files/working_papers/w32966/w32966.pdf

Anna Postol and Svitlana Tomko, "AI Overviews Research: How Google's AI Answers Vary Across Five States in the US," SE Ranking, accessed 20 September 2025, https://seranking.com/blog/ai-overviews-us-states-comparison-research/

Arooj Ahmed, "ChatGPT Usage Statistics: Numbers Behind Its Worldwide Growth and Reach (September, 2025)" Digital Information World, accessed May 20, 2025, https://www.digitalinformationworld.com/2025/05/chatgpt-stats-in-numbers-growth-usage-and-global-impact.html

Astrid Mager, Ov Cristian Norocel, and Richard Rogers, "Advancing search engine studies: The evolution of Google critique and intervention," Big Data & Society, 10 (2), 2023.

Axel G. Ekström, Guy Madison, Erik J. Olsson, and Melina Tsapos, "The search query filter bubble: effect of user ideology on political leaning of search results through query selection," Information, Communication & Society, 27 (5), 2023: 878-894.

Beatriz Botero Arcila, "Is it a Platform? Is it a Search Engine? It's Chat GPT! The European Liability Regime for Large Language Models," Journal of Free Speech Law, Vol. 3, Issue 2, https://ssrn.com/abstract=4539452, (2023).

Bernhard Rieder and Yarden Skop, “The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API,” Big Data & Society, 8(2), 2021: 20539517211046181

Brian Barrett, “Google AI Overviews Meaning,” Wired, accessed September 20, 2025, https://www.wired.com/story/google-ai-overviews-meaning/?sp=e823d799-2747-40c0-ac95-5a5f552fc28f.1747840680404

Cameron Pattison, Vance Ricks, and John Wihbey, “How AI-Driven Search May Reshape Democracy, Economics, and Human Agency,” Tech Policy Press, accessed September 20, 2025, https://www.techpolicy.press/how-ai-driven-search-may-reshape-democracy-economics-and-human-agency/.

Carolina Aguerre, Rikke Frank Jørgensen, Gry Hasselbalch, Frank Pasquale, Nathalie Smuha, Natalia Stanusch and Aimee van Wynsberghe (2023). “Generating AI: A Historical, Cultural, and Political Analysis of Generative Artificial Intelligence.” DataEthics.eu. https://dataethics.eu/generating-ai-a-historical-cultural-and-political-analysis-of-generative-artificial-intelligence/

Chandraveer Mathur, “Google's AI Mode rolls out nationwide with powerful new tools on the way,” Android Police, accessed May 20, 2025, https://www.androidpolice.com/google-ai-mode-rollout-new-tools-overviews/

Colette Stallbaumer, “Microsoft 365 Copilot: Built for the Era of Human-Agent Collaboration,” Microsoft. April 23, 2025, https://www.microsoft.com/en-us/microsoft-365/blog/2025/04/23/microsoft-365-copilot-built-for-the-era-of-human-agent-collaboration/

Cristiano Lima-Strong, “Google Dodges Breakup In Landmark Antitrust Ruling Over Its Search Engine,” Tech Policy Press, accessed September 20, 2025, https://www.techpolicy.press/google-dodges-breakup-in-landmark-antitrust-ruling-over-its-search-engine/

Council of the European Union, “General Approach,” November 25, 2022, accessed September 20, 2025, https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf/

Cynthia Kroet, “Zalando’s legal case to spark EU Commission online platform headache,” Euronews, accessed September 20, 2025, https://www.euronews.com/next/2025/03/06/zalandos-legal-case-to-spark-eu-commission-online-platform-headache

Daniel A Crane, “After Search Neutrality: Drawing a Line Between Promotion and Demotion,” I/S: J. L. & Pol'y for Info. Soc'y 9, no. 3 (2014): 397–406

Eamonn O'Raghallaigh, “The Key SEO Trend in 2025 – The Rise of Conversational AI Search,” Digital Strategy, accessed September 20, 2025, https://digitalstrategy.ie/insights/the-key-seo-trend-in-2025-the-rise-of-conversational-ai-search/

Eli Pariser, The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think (New York: Penguin, 2011)

Emillie de Keulenaar, “LLMs and the generation of moderate speech,” p. 5, accessed May 20, 2025, http://dx.doi.org/10.2139/ssrn.5250537.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜” in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (New York: Association for Computing Machinery, 2021), 610–623

Emma R. Murphy, Nadia Madkour, Deepika Raman, Kelsey Jackson, and Jesse Newman, “Survey of Search Engine Safeguards and their Applicability for AI,” 21

European Parliament, “Digital Services Act Application Timeline,” October 27, 2022, accessed September 20, 2025, https://www.europarl.europa.eu/RegData/etudes/ATAG/2022/739227/EPRS-AaG-739227-DSA-Application-timeline-FINAL.pdf

Esme Harrington and Mathias Vermeulen, “ “External researcher access to closed foundation models: State of the field and options for improvement,” The Mozilla Foundation, (2024).

EU Artificial Intelligence Act, “Recital 110,” Artificial Intelligence Act, accessed September 20, 2025, https://artificialintelligenceact.eu/recital/110/

EU Artificial Intelligence Act, “Recital 118,” Artificial Intelligence Act, accessed September 20, 2025, https://artificialintelligenceact.eu/recital/118/

European Commission, “Commission Compels Microsoft Provide Information,” European Commission, accessed September 20, 2025, https://digital-strategy.ec.europa.eu/en/news/commission-compels-microsoft-provide-information-under-digital-services-act-generative-ai-risks

European Commission, "Commission publishes guidelines under the DSA for the mitigation of systemic risks online for elections," accessed May 20, 2025, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1707

European Commission, "Commission Sends Requests Information Generative AI Risks," European Comission, accessed September 20, 2025, https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very

European Commission, "Contents Code GPAI," European Commission, accessed September 20, 2025, https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

Fabio Duarte, "Number of ChatGPT Users (October 2025)" Exploding Topics, accessed May 20, 2025, https://explodingtopics.com/blog/chatgpt-users

Gabriel Nicholas, "Grounding AI Policy: Towards Researcher Access to AI Usage Data," The Center for Democracy & Technology, accessed 20 September, 2025, https://cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/

Gemini, "About," Gemini.google, accessed May 20, 2025, https://gemini.google/about/

Google, "Terms Service Specific," Google Privacy & Terms, accessed September 20, 2025, https://policies.google.com/terms/service-specific

Haoyi Xiong et al., "When Search Engine Services Meet Large Language Models: Visions and Challenges," Arxiv, accessed September 20, 2025, arXiv:2407.00128.

Harry Booth, "Exclusive: 60 U.K. Lawmakers Accuse Google of Breaking AI Safety Pledge," The Times, accessed September 20, 2025, https://time.com/7313320/google-deepmind-gemini-ai-safety-pledge/

Lisa Eadicicco and Clare Duffy, "Google will not be forced to sell off Chrome or Android, judge rules in landmark antitrust ruling," CNN, accessed 20 September 2025, https://edition.cnn.com/2025/09/02/tech/google-antitrust-ruling-chrome-android

Ilan Strauss, Jangho Yang, Tim O'Reilly, Sruly Rosenblat, and Isobel Moure, "The Attribution Crisis in LLM Search Results," ArXiv, 2025, arXiv:2508.00838

Imran Rahman-Jones, "Meta AI searches made public – but do all its users realise?" BBC, accessed September 20, 2025, https://bbc.com/news/articles/c0573lj172jo

Jeff Horwitz, "Meta's AI rules have let bots hold 'sensual' chats with kids, offer false medical info," Reuters, accessed September 20, 2025, https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/

Jiawei Yao, Kai Ning, Zhiyuan Liu, Maosong Ning, and Ling Yuan, “LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples,” ArXiv, 2024, arXiv:2310.01469

Kashmir Hill and Dylan Freedman, “Chatbots Can Go Into a Delusional Spiral. Here's How It Happens,” The New York Times, accessed September 20, 2025, https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html

Kathleen A. Creel, “Transparency in Complex Computational Systems,” Philosophy of Science, 87 (4), 2020: 568–589

Kuai et al., 2024

Kyle Wiggers, “ChatGPT Isn’t the Only Chatbot that’s Gaining Users,” TechCrunch, accessed May 20, 2025, https://techcrunch.com/2025/04/01/chatgpt-isnt-the-only-chatbot-thats-gaining-users/

Lakera, “Lakera Guard Platform,” accessed September 20, 2025, https://platform.lakera.ai/

Lisa, Mays, “The consequences of search bias: how application of the essential facilities doctrine remedies Google's unrestricted monopoly on search in the United States and Europe,” George Washington Law Review, 83(2), 721–760, 2015

Luca Bertuzzi, “AI Act: MEPs want fundamental rights assessments, obligations for high-risk users,” Euractiv, accessed September 20, 2025, https://www.euractiv.com/news/ai-act-meps-want-fundamental-rights-assessments-obligations-for-high-risk-users/

Luca Bertuzzi, “ChatGPT Faces Possible Designation,” MLex, accessed September 20, 2025, https://www.mlex.com/mlex/articles/2332484/chatgpt-faces-possible-designation-as-a-systemic-platform-u

Marc Zao-Sanders, “How People Are Really Using Gen AI in 2025,” Harvard Business Review, accessed September 20, 2025, https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

Mario Haim, Andreas Graefe, and Hans-Bernd Brosius, "Burst of the Filter Bubble? Effects of personalization on the diversity of Google News," Digital Journalism, 6 (3), 2017: 330–343

Mark Gurman, “Apple Plans AI-Powered Web Search Tool for Siri to Rival OpenAI, Perplexity,” Bloomberg, accessed September 20, 2025, https://www.washingtonpost.com/technology/2020/10/06/amazon-apple-facebook-google-congress/

Mark MacCarthy, “The Privacy Challenges of Emerging Personalized AI Services,” Tech Policy Press, accessed September 20, 2025, https://www.techpolicy.press/the-privacy-challenges-of-emerging-personalized-ai-services/

Mathias Vermeulen and Laureline Lemoine, “From ChatGPT to Google’s Gemini,” LSE Media Blog, accessed September 20, 2025, https://blogs.lse.ac.uk/medialse/2024/02/12/from-chatgpt-to-googles-gemini-when-would-generative-ai-products-fall-within-the-scope-of-the-digital-services-act

Max Roslyakov, “SEO Market Stats (2024),” Xamsor, accessed 20 September 2025, https://xamsor.com/blog/seo-market-stats/

Maximilian Henning, “Danes Ask for Countries AI Act Simplification Wish Lists,” Euractiv, accessed September 20, 2025, https://www.euractiv.com/news/exclusive-danes-ask-for-countries-ai-act-simplification-wish-lists/.

Microsoft, “Introducing the New Bing. The AI-Powered Assistant for Your Search,” Feature & Tips, accessed 20 May 2025, https://www.microsoft.com/en-us/edge/features/the-new-bing

Michael Latzer, Katharina Hollnbuchner, Natascha Just, Florian Saurwein, “The economics of algorithmic selection on the Internet,” In: Johannes M. Bauer and Michael Latzer (ed.), Handbook on the Economics of the Internet p. 395-425 (Edward Elgar Publishing, 2016).

Michael Luca, Tim Wu, Sebastian Couvidat and Daniel Frank, “Does Google Content Degrade Google Search?” Experimental Evidence, Harvard Business School NOM Unit Working Paper No. 16-035, 2015, https://scholarship.law.columbia.edu/faculty_scholarship/1931

Microsoft, “Bing Systemic Risk Assessment Report,” accessed September 20, 2025, https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp//microsoft/final/en-us/microsoft-brand/documents/August-2024-Microsoft-Bing-Systemic-Risk-Assessment-Report-EU-Digital-Services-Act.pdf

Mistral AI, “Tech Details, Crunchbase, accessed September 20, 2025, https://www.crunchbase.com/organization/mistral-ai/technology

Natalia Stanusch, N., Buse Raziye Çetin, Salvatore Romano, Miazia Schueler, Meret Baumgartner, Bastian August, Alexandra Roșca. “LLMs, DSA, and AI Act: Introducing Methods and Approaches to Auditing LLMs Moderation across Languages and Interfaces in the Electoral Contexts.' In Rogers, R. (Ed) Content Moderation: A Cross-Platform Analysis, Amsterdam University Press. arXiv:2509.19890.

Nikos Smyrnaios, “Google as an Information Monopoly,” Contemporary French and Francophone Studies, 23(4), 442–446, 2019. DOI: 10.1080/17409292.2019.1718980

Océane Herrero, “Comment Apple apprend à son intelligence artificielle à s’adapter à l’ère Trump,” Politico, accessed September 20, 2025,

OpenAI , “Why Language Models Hallucinate,” OpenAI, accessed September 20, 2025, https://openai.com/index/why-language-models-hallucinate/

OpenAI, “ChatGPT. Overview,” OpenAI, accessed May 20,2025, https://openai.com/chatgpt/overview/

OpenAI, “GPT-5 System Card,” accessed September 20, 2025, https://cdn.openai.com/gpt-5-system-card.pdf

Paddy Leerssen, “Embedded GenAI on Social Media: Platform Law Meets AI law,” DSA Observatory, accessed September 20, 2025, https://dsa-observatory.eu/2024/10/16/1864/

Pascal Jürgens and Brigit Stark, “The Power of Default on Reddit: A General Model to Measure the Influence of Information Intermediaries,” Policy & Internet, 9: 395–419, 2017, DOI:10.1002/poi3.166

Patrick Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” 34th Conference on Neural Information Processing Systems, Article 793, 2020: 9459–9474, doi: 10.5555/3495724.3496517

Philipp Hacker and Atoosa Kasirzadeh and Lilian Edwards, “AI, Digital Platforms, and the New Systemic Risk,” https://ssrn.com/abstract=5475049

Philipp Hacker, Andreas Engel, and Marco Mauer, “Regulating Chat-GPT and other Large Generative AI Models. In 2023 ACM Conference on Fairness, Accountability, and Transparency” (FAccT ’23), June 12–15, 2023, ACM, New York, NY, USA. https://doi.org/10.1145/3593013.3594067

Richard Rogers, “Algorithmic probing: Prompting offensive Google results and their moderation,” Big Data & Society,10(1), 2023

Robert Epstein and Ronald E. Robertson, “The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections,” Proceedings of the National Academy of Sciences, 112 (33), 2015: E4512-E4521

Robert Epstein, Savannah Aries, Kally Grebbien, Alyssa M. Salcedo, and Vanessa R. Zankich, "The Search Suggestion Effect (SSE): How Autocomplete Search Suggestions Can Be Used to Impact Opinions and Votes," Computers in Human Behavior, 160C, 2024

Safiya Umoja Noble, Algorithms of Oppression: How Search Engines Reinforce Racism (New York: NYU Press, 2018)

Salvatore Romano, et al., "Chatbots: (S)Elected Moderation. Measuring the Moderation of Election-Related Content Across Chatbots, Languages and Electoral Contexts," AI Forensics, University of Amsterdam. https://aiforensics.org/work/chatbots-moderation (2024).

Salvatore Romano, et al., "Prompting Elections: The Reliability of Generative AI in the 2023 Swiss and German Elections," AI Forensics & Algorithm Watch (2023), https://aiforensics.org/work/bine-chat-elections

Sara Hooker, "On the Limitations of Compute Thresholds as a Governance Strategy," ArXiV preprint, arXiv:2407.05694v2

Sarthik Shah, Shantanu Neema, Rohan Singh Rajput, "Content Moderation Framework For The LLM-Based Recommendation Systems," International Journal of Computer Engineering and Information Technology, 14, pp. 104-117, (2024)

Shahan Ali Memon and Jevin D. West, "Search Engines Post-Chatgpt: How Generative Artificial Intelligence Could Make Search Less Reliable," Arxiv, accessed October 10, 2025, arXiv:2402.11707

Stuart A. Thompson, Teresa Mondría Terol, Kate Conger, and Dylan Freedman, "Elon Musk Grok Conservative Chatbot," The New York Times, accessed September 20, 2025, https://www.nytimes.com/2025/09/02/technology/elon-musk-grok-conservative-chatbot.html

Tamar Sharon, "Towards a theory of justice for the digital age. In defence of sphere and value pluralism," Inaugural lecture, available at: https://repository.ubn.ru.nl/bitstream/handle/2066/300467/300467.pdf?sequence=1.

Tarleton Gillespie, "Regulation of and by platforms," in SAGE Handbook of Social Media, edited by Jean Burgess, Thomas Poell, and Alice Marwick, Sage, (2017).

Tarleton Gillespie, "The Relevance of Algorithms," in Tarleton Gillespie, Pablo J. Boczkowski, Kirsten A. Foot (eds) Media Technologies: Essays on Communication, Materiality, and Society (Cambridge, MA: The MIT Press, 2014), 168

The Digital Services Act was published in the Official Journal on October 27, 2022

Tiago Bianchi, “Global market share of lead consolidated new forms of powering desktop search engines 2015–2025,” Statista, accessed 20 September, 2025, https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Tiago Bianchi, “Number of Adults in the United States Using Generative Artificial Intelligence (AI) First for Online Search in 2024 and 2028,” Statista, May 28, 2025, https://www.statista.com/statistics/1454204/united-states-generative-ai-primary-usage-online-search/

Tony Romm et al., “House investigation faults Amazon, Apple, Facebook and Google for engaging in anti-competitive monopoly tactics,” The Washington Post, accessed September 20, 2025, https://www.washingtonpost.com/technology/2020/10/06/amazon-apple-facebook-google-congress/

"Tracking Large Scale AI Models," Epoch AI, accessed September 20, 2025, https://epoch.ai/blog/tracking-large-scale-ai-models#benchmarks-and-repositories

X.AI, “Bringing Grok to Everyone,” X.AI, accessed May 20, 2025, https://x.ai/news/grok-1212

Yue Zhang et al., “Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models,” ArXiv, arxiv.org/pdf/2309.01219, 2024, 1

Ziyou Yan, “Improving Recommendation Systems & Search in the Age of LLMs,” eugeneyan.com, accessed 20 May 2025, https://eugeneyan.com/writing/recsys-llm/

AI FORENSICS

November 2025

R. Buse Çetin Natalia Stanusch Marc Faddoul

From “Googling” to “Asking ChatGPT”:

Governing AI Search