How Rules for Publicly Available Data Are Shaping the Future of AI

By Daniel Castro | March 13, 2026

Whether nations permit AI systems to learn from publicly available information will shape global leadership in artificial intelligence (AI). Restrictive rules on the use of public web data risk shifting AI development to more permissive jurisdictions, undermining a country's ability to build, deploy, and benefit from next-generation AI systems. A more effective approach emphasizes technical opt-outs, transparency, and safeguards that prevent harmful outputs. Policies that preserve responsible access to the digital commons will better support the next generation of AI capabilities and economic growth.

INTRODUCTION

Artificial intelligence (AI) systems learn by analyzing vast quantities of digital information. As governments debate how to regulate AI, a central question has emerged: Should developers be allowed to train models on information that is publicly available on the Internet, even when that information contains personal data?

The answer will shape not only privacy protections but also the future trajectory of AI development. Publicly accessible websites, open databases, government records, and other online resources form a critical pool of knowledge that AI systems rely on to understand language, reason about the world, and verify information. At the same time, the ability of AI systems to analyze this information at scale raises legitimate questions about how AI systems should use and protect personal data.

Different jurisdictions have begun to approach these questions in different ways. The United States generally treats publicly accessible web data as available for automated collection unless site owners impose technical

barriers. The European Union, by contrast, places broader restrictions on how organizations may process personal data—even when that information appears on public websites. As AI capabilities advance and agentic systems begin interacting with information and services across the Internet, these policy differences increasingly will influence where AI development occurs, and which countries will capture the economic benefits of AI adoption.

Policymakers can protect individuals while preserving the open information ecosystem that supports innovation. This approach can be grounded in three key principles:

1. **Focus on outputs rather than training inputs.** Address harmful uses of AI systems—such as revealing sensitive personal information—instead of restricting the collection of publicly available data for model training.

2. **Encourage transparency norms for autonomous AI agents.** Promote voluntary industry practices for AI developers to help people understand when they are interacting with automated systems, while allowing flexibility for evolving uses of agentic AI.

3. **Create a safe harbor for responsible use of publicly available data.** Provide legal certainty for developers that respect machine-readable opt-out signals from websites and that use automated tools to filter sensitive personal information during data preparation.

The Internet has long served as a shared source of public knowledge. In the AI era, it has become a foundational input for building systems that can reason, retrieve information, and interact effectively with the world. Policymakers in the United States, the European Union, and everywhere else that aspires to be at the forefront of AI development and adoption should ensure developers can continue using it that way.

WHY PUBLIC DATA MATTERS FOR AI

AI systems learn by analyzing vast amounts of data. Developers train AI models by exposing them to large collections of text, images, code, and other information so the models can identify patterns in language and other types of content. During training, models break text into small units called tokens—typically words or parts of words—and learn statistical relationships among them. By analyzing trillions of tokens, these systems learn how to generate coherent text, answer questions, write code, and perform many other tasks.

Much of the data used to train these systems comes from the public Internet. Websites, open databases, government publications, scientific research, and collaborative resources such as Wikipedia together form a kind of digital commons—a shared pool of knowledge that reflects the

diversity of human language, culture, and expertise. Developers combine this publicly available data with licensed datasets, proprietary materials, and synthetic data to train general-purpose AI models.

Publicly available data plays a particularly important role in enabling general-purpose reasoning. Foundation models—the large AI systems that power modern chatbots and other applications—require enormous volumes of information to develop broad capabilities. Recent models have trained on trillions of tokens.[1] If developers relied only on licensed datasets such as news archives or academic journals, the resulting models would reflect a narrower slice of human knowledge. They would struggle to understand informal language, regional dialects, niche technical topics, and everyday cultural references that appear widely across the open web. In other words, they would become brittle systems that perform well on specialized content but poorly on the messy diversity of real-world communication.

Publicly available datasets also serve a second critical set of functions: verification and benchmarking. Reliable AI systems ground their outputs in accurate information. Open resources such as government statistics, public records, open scientific databases, and collaboratively curated knowledge bases provide widely accepted reference points—a collective "ground truth." Developers use these datasets throughout the AI lifecycle: during initial training, during fine-tuning to improve accuracy, and during evaluation to measure whether models produce correct answers.

These datasets also play an increasing role in the development of agentic AI systems. Unlike traditional chatbots that respond to a single prompt, agentic systems can plan tasks, retrieve information, and take actions across multiple steps. To operate safely and reliably, they must consult trustworthy external information sources. Publicly available datasets and open web resources therefore act as reference material that agents can query to verify claims, retrieve factual information, and check the accuracy of their outputs.

For these reasons, access to publicly available Internet data has become a foundational input for AI development. Differences in data accessibility will likely shape global AI competition. In 2025, U.S.-based organizations produced 40 notable foundation models, while organizations in China produced 15 and those in the European Union produced only 3—an emerging "model gap" that may reflect differences in access to training data, among other factors.[2]

At the same time, the public availability of data raises legitimate concerns about how personal information circulates online and how AI systems might use it. Policymakers therefore must confront a central tension of the AI era: The same publicly available data that enables powerful and broadly useful AI systems also can create new privacy risks.

PRIVACY CONCERNS AND THE PUBLIC INTERNET

Debates about AI and publicly available data often begin with a basic reality: Information that is publicly available online can be reused and analyzed at scales or in contexts that the original publisher may not have anticipated. The Internet contains vast amounts of personal information that individuals, organizations, and governments publish for specific purposes—professional networking, civic transparency, research, or communication. As AI systems analyze this information in new ways, policymakers should consider how existing expectations about public information apply in the AI era, as well as how privacy norms might change over time.

Much of the public Internet contains personally identifiable information (PII)—data that can identify a specific individual. Social media profiles may reveal a person's age, employer, or hometown. Professional websites often list employment histories and educational backgrounds. Government filings can disclose property ownership, campaign contributions, or business registrations. Because this information appears on publicly accessible webpages, search engines have long crawled and indexed it to help users locate relevant information online.

AI changes the scale and manner in which systems can process this information. Traditional search engines direct users to existing webpages, while AI systems can analyze large collections of documents and generate summaries or answers that synthesize information from multiple sources. This capability can improve productivity and research, but it also raises questions about how easily personal information can be aggregated or surfaced through automated systems.

Some observers also point to the persistence of digital information. Human interactions historically included many fleeting moments—casual conversations, temporary records, or informal communications—most of which faded over time. The Internet has changed this dynamic by creating durable digital archives. Similar concerns emerged when earlier technologies spread widely, including photography, recorded media, and online search engines.[3] In each case, society adapted to technologies that preserved or made information easier to retrieve.

AI introduces related questions in a new technical context. When developers train models, they incorporate patterns from large datasets into billions or trillions of internal parameters known as model weights. Because these systems learn statistical relationships rather than storing documents directly, removing the influence of specific information after training can be difficult. Researchers continue to explore technical approaches such as dataset filtering, model editing, and improved training methods to address these challenges.

Finally, discussions about publicly available data and AI often include broader societal considerations. Large-scale automated analysis can surface patterns or connections that would have been difficult to identify manually. In many contexts, this capability creates clear benefits—for example, by helping researchers analyze scientific literature or enabling journalists to examine large datasets of public records. At the same time, it raises questions about how to balance innovation with reasonable expectations about how personal information published online may be reused.

These tensions are not new, but AI brings them into sharper focus. As policymakers evaluate the role of publicly available data in AI development, they should first consider how existing legal and regulatory frameworks apply to the collection and use of publicly accessible information and address both individual interests and the public interest in technological progress.

US AND EU REGULATORY DIVERGENCES

Regulatory approaches to publicly available data have long diverged across jurisdictions, but the rapid rise of AI development has made those differences far more consequential. The United States and the European Union—two of the world's largest technology markets—have adopted distinct legal frameworks governing how developers may collect and use publicly available information online. This once reflected philosophical differences about privacy and openness, but now it produces tangible operational consequences for AI developers, shaping where companies train models, launch products, and invest in new research.

Table 1: The crawling landscape, US vs. EU

Dimension	United States	European Union
Overall position	Permissive (publicly available data may be freely used)	Restricted (rights follow the data)
Public data	Excluded from most privacy rights	Subject to GDPR protections
Facial data	Legal, except for specific state restrictions	Banned (AI Act, Article 5)
Machine-readable opt-outs	Voluntary compliance	Mandatory compliance (Copyright Directive, Article 4)
Untraining	Not required	Potentially required (GDPR, Article 17)

The United States: A "Gates Up" Approach

U.S. law generally treats publicly accessible Internet data as available for automated collection unless website operators impose clear technical barriers.

One key legal framework is the Computer Fraud and Abuse Act (CFAA), the primary federal statute governing unauthorized access to computer systems.[4] In *Van Buren v. United States* (2021), the U.S. Supreme Court held that the CFAA applies when a person accesses areas of a computer system that are off limits, not when they access information that is otherwise available to them.[5] Federal courts have applied this reasoning in web-scraping disputes. For example, in *hiQ Labs v. LinkedIn*, the U.S. Court of Appeals for the Ninth Circuit concluded that scraping publicly accessible LinkedIn profile data likely does not violate the CFAA because the information does not sit behind an authentication barrier.[6] In other words, if a website does not place data behind a password or similar "digital gate," courts have generally treated that information as accessible. This framework has allowed AI developers to collect large volumes of publicly available web data for training models.

Copyright law also places some limits on how developers may collect online data. The Digital Millennium Copyright Act (DMCA) prohibits circumventing technological protection measures that control access to copyrighted works.[7] As a result, developers generally cannot bypass paywalls, authentication systems, or other access controls in order to obtain protected content for training datasets. While this constraint arises from copyright law rather than privacy law, it illustrates how U.S. policy distinguishes between information that is publicly accessible and information that is deliberately placed behind technical access barriers.

At the same time, state privacy laws increasingly regulate how companies disclose and manage data practices. California has emerged as a leading example. Recent legislation such as AB 2013 (2024) requires developers of generative AI systems to publish high-level summaries of the datasets used for training, including whether datasets include personal information.[8]

Similarly, the California Delete Act (SB 362) gives consumers the right to request deletion of personal information held by data brokers.[9] However, California privacy laws—including the California Consumer Privacy Act (CCPA)—continue to exempt "publicly available information" from many of their most restrictive provisions.[10]

Some state laws nonetheless impose clear boundaries. For example, the Illinois Biometric Information Privacy Act (BIPA) strictly regulates the collection of biometric identifiers such as facial geometry or voiceprints.[11] Courts have interpreted the statute to prohibit companies from collecting biometric data for identification purposes without informed consent,

making it one of the most consequential privacy laws affecting AI training and facial recognition technologies in the United States.

Taken together, U.S. policy has largely allowed developers to access publicly available web data, even when that data contains personal information.

The European Union: A Rights-Based Framework

The European Union approaches publicly available data through a broader privacy framework centered on fundamental rights. The General Data Protection Regulation (GDPR) governs how organizations process personal data, including information that appears on publicly accessible websites.

Under Article 6 of the GDPR, organizations must identify a lawful basis for processing personal data.[12] One such basis is "legitimate interest," which requires organizations to demonstrate that their interests outweigh the privacy interests of individuals. The regulation also grants individuals rights over their personal data, including the right to erasure under Article 17, which allows individuals to request deletion of personal data in certain circumstances.[13] Some European regulators have indicated that these rights may apply to AI developers that train models on personal data scraped from the web, raising questions about how such requests could be implemented once information has been incorporated into a model's parameters.[14]

Copyright law also intersects with the use of publicly available data for AI training. Article 4 of the EU's Directive on Copyright in the Digital Single Market allows text and data mining of publicly accessible content but permits rights holders to opt out by reserving their rights through machine-readable means.[15] As a result, AI developers operating in Europe must navigate both copyright opt-outs and data-protection rules when collecting information from the open web.

In practice, this standard introduces legal uncertainty for large-scale web scraping. Regulators must assess whether collecting publicly available personal data to train AI systems satisfies the legitimate interest test and whether the scope of the collection remains proportionate. European data-protection authorities have increasingly scrutinized this balance as generative AI systems scale.

The EU has also introduced AI-specific legislation. The Artificial Intelligence Act, adopted in 2024, includes targeted restrictions on certain data-collection practices. Article 5 prohibits the creation or expansion of facial-recognition databases through the untargeted scraping of facial images from the Internet, reflecting concerns about mass biometric identification.[16]

These requirements interact with broader compliance obligations under the GDPR, including documentation, risk assessments, and data-protection

safeguards. For start-ups and smaller developers, these obligations can translate into substantial administrative costs. Several studies and industry surveys have suggested that regulatory complexity contributes to concerns about the European Union's competitiveness in AI development and may influence where some companies choose to locate research and infrastructure.[17]

MARKET-LED GOVERNANCE

Alongside regulatory debates, the Internet ecosystem has begun developing technical mechanisms that allow publishers and AI developers to coordinate how automated systems access and use online content. Rather than forcing websites to choose between fully blocking or fully allowing AI crawlers, these tools support a more "negotiated web" in which site owners can express granular preferences and developers can program their systems to respect those signals automatically.

Many of these mechanisms build on long-standing Internet standards while introducing new protocols designed specifically for AI systems.

Advanced Discovery and Curation

The most widely used mechanism for controlling automated access remains robots.txt, a standard introduced in the 1990s that allows website operators to specify which automated agents may crawl particular pages or directories.[18] Search engines and most major AI developers honor these directives as part of industry norms for responsible crawling.

Related discovery tools help crawlers locate relevant information efficiently. For example, XML sitemaps provide structured lists of a site's pages so automated systems can index content more reliably. More recently, some publishers have begun experimenting with LLMs.txt, an emerging convention in which a website provides a curated, machine-readable summary of its content—often written in Markdown, a simple plain-text formatting language commonly used to structure documentation on the web—to help large language models access authoritative information about the site.[19] Companies such as Stripe and Vercel have adopted this approach as a way to guide AI systems toward high-quality explanations of their documentation or services while reducing the amount of irrelevant text that models must process.[20]

Although LLMs.txt is not yet an official Internet standard, it reflects a broader trend toward curated machine-readable content designed specifically for AI systems.

The Emergence of AI Preferences

Researchers and standards bodies have begun exploring more granular ways for websites to communicate how AI systems may use their content.

One emerging concept is the development of standardized AI preference signals that extend beyond the binary allow-or-block structure of robots.txt.

Within the Internet Engineering Task Force (IETF)—the organization responsible for many core Internet standards—researchers have proposed mechanisms that would allow websites to express more specific policies for automated systems.[21] These could include distinctions between different uses of data, such as permitting indexing for search while prohibiting training for AI.

In practice, such signals could be delivered through machine-readable metadata, HTTP headers, or configuration files at the root of a website. If widely adopted, they would allow automated systems to interpret and comply with site-specific policies programmatically, making it easier for developers to respect publishers' preferences at scale.

Cryptographic Trust and Bot Authentication

Another challenge in managing automated access involves verifying the identity of crawlers. Website operators often rely on user-agent strings to identify bots, but these identifiers can be easily spoofed by malicious actors.

To address this issue, researchers have proposed cryptographic bot authentication systems that allow automated agents to verify their identity when accessing websites. One approach builds on HTTP message signatures, a developing standard that enables a client to cryptographically sign requests so that the receiving server can confirm the sender's identity.[22]

In this model, a crawler operated by a research institution or technology company could attach a verifiable signature to each request. Websites could then confirm that a bot claiming to belong to a particular organization genuinely does. If a crawler violated a site's policies, the signature would provide a clear audit trail linking the activity to a specific operator.

Monetization and Programmatic Licensing

The emerging "negotiated web" also creates new opportunities for commercial data licensing. One mechanism relies on the HTTP 402 "Payment Required" status code, which signals that access to a resource requires payment.[23] Although the code has historically seen limited use, some platforms are exploring ways to employ it as part of automated licensing frameworks.

For example, web infrastructure providers like Cloudflare have begun experimenting with systems that allow publishers to signal to AI crawlers that certain content requires a paid license. In principle, a crawler encountering such a signal could automatically initiate a licensing

transaction or redirect to a payment mechanism. While these systems remain early in development, they illustrate how technical standards could support programmatic marketplaces for data access.

Privacy Safeguards During Data Pre-Processing

AI developers also implement technical safeguards during the data preparation stage before training models. Many organizations apply automated filtering tools to identify and remove sensitive personal information from datasets. One widely used example is Microsoft Presidio, an open-source toolkit designed to detect and anonymize PII such as Social Security numbers, phone numbers, addresses, and credit card details in unstructured text.[24] These systems allow developers to mask or remove sensitive data during ingestion, reducing privacy risks without requiring publicly available content to be completely excluded from training datasets.

Together, these tools illustrate how technical standards and market mechanisms can complement legal frameworks in governing AI data practices. Instead of relying solely on regulation, voluntary efforts from the Internet ecosystem increasingly supports machine-readable signals, authenticated crawlers, and automated licensing systems that allow publishers and developers to coordinate how AI systems access and use online information.

HOW AI AGENTS RAISE THE STAKES

Most of the debate over publicly available data and AI has focused on web crawling—the automated collection of information from publicly accessible websites. But the rapid development of agentic AI systems is expanding the scope of the issue. Unlike traditional AI models that simply generate responses to user prompts, agentic systems can plan tasks, access external tools, retrieve information, and execute actions on behalf of users. These capabilities shift the focus of privacy discussions from static datasets toward how AI systems interact with information and permissions in real time.

From Crawling to Delegated Access

A defining feature of many agentic systems is their ability to act on a user's behalf with their existing accounts and permissions. Users increasingly authorize AI assistants to access services such as email, calendars, or professional platforms in order to summarize messages, schedule meetings, or gather information. This delegated access allows agents to operate using credentials or application programming interfaces (APIs) tied to the user's identity.

This shift complicates earlier legal frameworks designed around public web crawling. When a developer scrapes publicly accessible websites, the information typically does not sit behind an authentication barrier. Agentic systems, by contrast, may interact with information that is technically

private but legitimately accessible to the user. The result is a new category of "private-but-available" data—information that is not publicly accessible on the open web but that an agent can retrieve because it operates with the user's authorization.

Security researchers have also highlighted a related vulnerability known as indirect prompt injection.[25] In this scenario, a malicious actor embeds instructions within a webpage, document, or email that an AI agent later processes. Because the agent has access to a user's accounts or files, the hidden instructions may attempt to manipulate the agent into revealing sensitive information or performing unintended actions. Researchers describe this dynamic as a version of the "confused deputy" problem, a computer security concept in which a system with legitimate privileges is tricked into misusing them.[26]

These risks illustrate how agentic systems shift privacy concerns from data at rest—the information stored in datasets—to permissions in motion, where the key issue becomes how AI systems exercise the access rights delegated to them.

Proactive Outreach and Identity Transparency

Agentic AI systems can also initiate interactions rather than simply responding to user prompts. For example, an agent might send messages to coordinate meetings, contact businesses to gather information, or place automated phone calls to complete routine tasks.

These capabilities introduce questions about identity transparency in digital communications. When an AI system contacts a person, that individual may not always know whether they are interacting with a human or a machine. Several jurisdictions and industry groups have begun exploring disclosure requirements for AI-generated communications to address this challenge. For example, Article 50 of the European Union's AI Act requires providers to design AI systems so that individuals are informed when they are interacting with an AI system, particularly in the case of chatbots and voice assistants, unless the automated nature of the interaction is obvious from the context.[27] Similarly, California's Companion Chatbot Law (SB 243), which went into effect in 2026, requires operators of chatbots to disclose that the system is not a human if a reasonable person could otherwise be misled.[28] Other jurisdictions have taken a more targeted approach. Utah's Artificial Intelligence Policy Act, enacted in 2024, requires businesses to disclose when consumers are interacting with generative AI in certain regulated contexts, such as health care or legal services.[29]

Clear disclosure norms can help ensure automated outreach does not create confusion in everyday communications. Individuals may share information or respond differently depending on whether they believe they are interacting with a human representative or an automated system.

Connecting Disparate Data

Agentic systems can also combine information from multiple sources in ways that earlier technologies made more difficult. A single agent may have access to a user's email, calendar, professional profiles, and publicly available web information in order to complete tasks efficiently.

This capability enables useful services—for example, coordinating travel plans or organizing professional contacts—but it also highlights a broader data synthesis challenge. When systems combine information across platforms, they can reveal connections between datasets that were originally separate. In some contexts, this could allow agents to assemble detailed profiles of individuals using information drawn from a mixture of public and semi-private sources.

Agent-to-Agent Communication

A final emerging trend involves machine-to-machine interaction. Increasingly, AI agents communicate directly with other automated systems. For example, a user's scheduling assistant may coordinate with a restaurant's reservation system or a travel platform's booking agent.

These interactions often occur through APIs or automated negotiation protocols and can take place at machine speed. Because the exchange happens between software systems rather than through a human interface, users may have limited visibility into exactly what information their agents share during these interactions. In practice, this could include routine details such as location data, preferences, or scheduling information.

Agentic AI therefore expands the scope of the public-data debate. Traditional discussions have focused on whether developers could collect publicly available information from the web. Agentic systems raise a broader question: How should AI systems use access and permissions to personal information once they begin acting on behalf of users? As policymakers consider rules to govern AI data practices, they should account not only for training datasets but also for the increasingly dynamic ways AI systems interact with information across the digital ecosystem.

POLICY RECOMMENDATIONS

Policymakers should avoid regulatory approaches that inadvertently restrict access to the publicly available information that underpins modern digital innovation. The goal should not be to tightly control how developers collect publicly available data, but rather to ensure that AI systems are deployed responsibly and transparently. A light-touch framework can protect individuals while preserving the open information ecosystem that enables AI research, competition, and economic growth.

Three principles can guide such an approach.

1. Focus on Harmful Outputs Rather Than Training Inputs

Many policy proposals attempt to regulate AI development by dictating which data developers may or may not use to train models. In practice, this approach is both difficult to enforce and technologically misguided. Modern AI models do not store documents or databases in a retrievable form; instead, they learn statistical relationships across billions or trillions of parameters. Once training occurs, it becomes extremely difficult to isolate or remove the influence of any single data point—a process sometimes described as "machine unlearning," which remains an experimental research area rather than a practical compliance tool.[30]

Rather than attempting to police the entire training pipeline, policymakers should focus on preventing harmful outputs. Developers already implement safeguards that reduce the likelihood that models will generate or expose sensitive personal information. Governments can reinforce these practices by setting clear expectations around output safety—such as prohibiting systems from disclosing non-public personal information—while allowing developers flexibility in how they design and train their models. Existing legal frameworks already provide enforcement tools in many cases: If an AI system improperly exposes protected personal data, the resulting disclosure should be treated like any other unauthorized data disclosure under existing privacy or consumer-protection laws.

This approach addresses real risks without undermining the broad access to public information that allows AI systems to understand language, culture, and real-world knowledge.

2. Encourage Transparency Norms for Autonomous AI Agents

As AI systems evolve from passive tools into agents capable of sending messages, making calls, or gathering information, transparency will become increasingly important for maintaining trust in digital communication. Rather than imposing rigid disclosure mandates, policymakers should encourage the development of industry norms and best practices that promote transparency when AI systems interact directly with people. In many contexts—such as customer service, scheduling, or routine information requests—automated interactions may already be widely accepted. In others, such as when receiving health care or legal services, individuals may reasonably expect to know whether they are communicating with a human or an AI system.

Flexible transparency standards would allow these expectations to evolve as the technology matures. Companies could adopt practices such as identifying automated agents in certain contexts, labeling communications as AI-generated, human-generated, or AI-assisted, and offering users ways to request or verify human involvement when appropriate.

A norms-based approach preserves room for innovation while encouraging responsible deployment of agentic AI systems and maintaining trust in digital communications.

3. Create a Safe Harbor for Responsible Use of Publicly Available Data

Finally, policymakers—particularly in middle-power economies seeking to grow their AI sectors—should provide legal certainty for developers that rely on publicly available data. Access to the open web remains essential for building accurate, broadly capable AI systems. When regulations restrict the use of publicly available data for training, developers can shift model development to jurisdictions with more permissive rules, leaving more restrictive countries dependent on AI systems built abroad.

One solution would be to establish a safe harbor for developers that follow widely recognized industry practices when collecting and using public web data. To qualify, developers could be required to:

- respect machine-readable opt-out signals from website operators, such as directives in robots.txt or similar emerging standards;
- use automated tools to detect and filter sensitive personal information before incorporating data into training pipelines; and
- provide high-level transparency about the types of data used to train their systems.

A safe harbor framework would reward responsible practices while avoiding rigid rules that quickly become obsolete as AI technology evolves. It would also encourage the development of technical standards—such as machine-readable permissions and authenticated crawlers—that allow publishers and developers to coordinate data use efficiently.

Taken together, these principles offer a pragmatic path forward. By focusing on harmful outputs, promoting transparency in automated interactions, and providing legal certainty for responsible uses of publicly available data, policymakers can protect individuals without closing off the open information environment that has long powered innovation on the Internet.

CONCLUSION

Access to publicly available data will play a decisive role in determining which countries lead in the development of advanced AI systems. The open web functions as a shared knowledge infrastructure for modern machine learning, enabling models to learn language, verify facts, and interact effectively with the real world. Policies that unnecessarily restrict access to this digital commons risk weakening the foundations on which future AI innovation depends.

A forward-looking governance framework therefore should protect individuals while preserving the open information ecosystem that has long powered technological progress. Jurisdictions that strike this balance—supporting responsible data use, transparency, and technical safeguards without closing off access to public information—will be best positioned to lead in the next generation of AI development.

REFERENCES

1. Abhimanyu Dubey et al., “The Llama 3 Herd of Models,” arXiv, July 31, 2024 (v3 updated Nov 2024), https://arxiv.org/abs/2407.21783.
2. Yolanda Gil and Raymond Perrault, “The AI Index 2025 Annual Report,” Stanford Institute for Human-Centered AI (HAI), April 2025, https://hai.stanford.edu/ai-index/2025-ai-index-report.
3. Daniel Castro and Alan McQuinn, “The Privacy Panic Cycle: A Guide to Public Fears About New Technologies,” Information Technology and Innovation Foundation (ITIF), September 2015, https://www2.itif.org/2015-privacy-panic.pdf.
4. U.S. Department of Justice, Justice Manual, § 9-48.000, “Computer Fraud and Abuse Act,” updated May 2022, https://www.justice.gov/jm/jm-9-48000-computer-fraud.
5. Supreme Court of the United States, *Van Buren v. United States*, 593 U.S. 374 (2021), June 3, 2021, https://www.supremecourt.gov/opinions/20pdf/19-783_k53l.pdf.
6. United States Court of Appeals for the Ninth Circuit, *hiQ Labs, Inc. v. LinkedIn Corp.*, 31 F.4th 1153 (9th Cir. 2022), April 18, 2022, https://cdn.ca9.uscourts.gov/datastore/opinions/2022/04/18/17-16783.pdf.
7. U.S. Copyright Office, “The Digital Millennium Copyright Act of 1998: U.S. Copyright Office Summary,” December 1998, https://www.copyright.gov/legislation/dmca.pdf.
8. California Assembly Bill 2013, “Generative Artificial Intelligence: Training Data Transparency,” 2023-2024 Regular Session, September 28, 2024, https://legiscan.com/CA/text/AB2013/id/3023192.
9. California State Legislature, “Senate Bill No. 362: California Data Broker Act,” October 10, 2023, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB362.
10. California Civil Code, div. 3, pt. 4, tit. 1.81.5, “California Consumer Privacy Act of 2018,” accessed March 8, 2026, https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.
11. Illinois Senate Bill 2979, “Biometric Information Privacy Act: Section 20 Liability,” 103rd General Assembly, August 2, 2024, https://legiscan.com/IL/text/SB2979/id/2908858.
12. European Parliament and Council of the European Union, “General Data Protection Regulation (GDPR), Article 6: Lawfulness of processing,” May 25, 2018, https://gdpr-info.eu/art-6-gdpr/.
13. European Parliament and Council of the European Union, “General Data Protection Regulation (GDPR), Article 17: Right to Erasure ('Right to be Forgotten'),” May 24, 2016, https://gdpr-info.eu/art-17-gdpr/.
14. CNIL, “Ensuring and Facilitating the Exercise of Data Subjects' Rights,” January 5, 2026, https://www.cnil.fr/en/ensuring-and-facilitating-exercise-data-subjects-rights.
15. European Parliament and Council of the European Union, “Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019

on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC," Official Journal of the European Union, L 130, May 17, 2019, https://eur-lex.europa.eu/eli/dir/2019/790/oj.

16. European Commission, "AI Act, Article 5: Prohibited Artificial Intelligence Practices," August 1, 2024, https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-5.

17. European Commission, "The Draghi Report: The Future of European Competitiveness," September 9, 2024, https://commission.europa.eu/topics/competitiveness/draghi-report_en; AppliedAI, "AI Act Impact Survey: Understanding the Impact of the AI Act on the European AI Ecosystem," accessed March 10, 2026, https://www.appliedai.de/en/ai-resources/white-papers/ai-act-impact-survey/; ACT | The App Association, "The Hidden Cost of AI Regulations: A Survey of EU, UK, and U.S. Companies," October, 2025, https://actonline.org/the-hidden-cost-of-ai-regulations-a-survey-of-eu-uk-and-u-s-companies/.

18. Martijn Koster, Gary Illyes, Henner Zeller, and Lizzi Sassman, "Robots Exclusion Protocol," IETF RFC 9309, September 2022, https://datatracker.ietf.org/doc/html/rfc9309.

19. "llms.txt: A Proposed Standard for LLM-Friendly Content," accessed March 10, 2026, https://llmstxt.org/.

20. Gertjan De Wilde, "Stripe's llms.txt Has an 'Instructions' Section—That's a Bigger Deal Than It Sounds," Dev.to, March 4, 2026, https://dev.to/apideck/stripes-llmstxt-has-an-instructions-section-thats-a-bigger-deal-than-it-sounds-8ad.

21. Suresh Krishnan and Mark Nottingham, "Progress on AI Preferences," IETF Blog, Internet Engineering Task Force, February 25, 2025, https://www.ietf.org/blog/ai-pref-progress/.

22. Mari Galicer, Akshat Mahajan, Gauri Baraskar, and Helen Du, "Message Signatures Are Now Part of Our Verified Bots Program, Simplifying Bot Authentication," The Cloudflare Blog, July 1, 2025, https://blog.cloudflare.com/verified-bots-with-cryptography/.

23. Will Allen and Simon Newton, "Introducing Pay Per Crawl: Enabling Content Owners to Charge AI Crawlers for Access," *The Cloudflare Blog*, July 1, 2025, https://blog.cloudflare.com/introducing-pay-per-crawl/.

24. Microsoft, "Presidio: Data Protection and De-identification SDK," Microsoft Open Source, accessed March 10, 2026, https://microsoft.github.io/presidio/.

25. John Gamble, "Indirect Prompt Injection Attacks: A Lurking Risk to AI Systems," *CrowdStrike Blog*, December 4, 2025, https://www.crowdstrike.com/en-us/blog/indirect-prompt-injection-attacks-hidden-ai-risks/.

26. AWS, "The Confused Deputy Problem," IAM User Guide, accessed March 10, 2026, https://docs.aws.amazon.com/IAM/latest/UserGuide/confused-deputy.html.

27. European Parliament and Council of the European Union, "AI Act, Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems," August 1, 2024, https://artificialintelligenceact.eu/article/50/.

28. California Senate Bill 243, “Companion Chatbots,” 2025-2026 Regular Session, October 13, 2025, https://legiscan.com/CA/bill/SB243/2025.

29. Utah Senate, “S.B. 149: Artificial Intelligence Amendments,” 2024 General Session, March 13, 2024, https://le.utah.gov/~2024/bills/static/SB0149.html.

30. Ken Ziyu Liu, “Machine Unlearning in 2024,” Stanford Computer Science, May 2024, https://ai.stanford.edu/~kzliu/blog/unlearning/; A. Feder Cooper et al., “Machine Unlearning Doesn’t Do What You Think: Lessons for Generative AI Policy and Research,” arXiv, December 2024 (v2 updated November 2025), https://arxiv.org/abs/2412.06966.

ABOUT THE AUTHOR

Daniel Castro is vice president of ITIF and director of ITIF's Center for Data Innovation. He has a B.S. in foreign service from Georgetown University and an M.S. in information security technology and management from Carnegie Mellon University.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation studies the intersection of data, technology, and public policy. With staff in Washington, London, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

Contact: info@datainnovation.org

datainnovation.org