



Artificial Intelligence and Open Source Data and Software: Contrasting Perspectives, Legal Risks, and Observations

What You Need to Know

Key takeaway #1

Firms involved in open source data and AI need to understand their obligations under new laws and regulations. For example, the European Union's Cyber Resilience Act (CRA) creates specific obligations for open source software in its requirements for software manufacturers, importers, and distributors.

Key takeaway #2

U.S. regulators appear to be taking a cautious approach to the use of open source data and software, focusing on the lack of due diligence around updates to and control over open source software and potential risks in cyber and national security.

Key takeaway #3

Many intergovernmental organizations and open source advocates continue to push for data democratization, transparency, and public-private collaboration, especially in the context of AI governance and development.

Client Alert | 8 min read | 07.23.25

Open source data and software play a foundational role in software development, artificial intelligence (AI), education, and research. Open source AI refers to systems where the source code, model parameters, and related components are freely available for anyone to use, study, modify, and distribute.

Governments, international organizations, civil societies, and industry communities are increasingly at odds over how open-source data and software should be used, governed, and protected. For example, open source AI became part of U.S.-China geopolitical competition after the **Chinese firm DeepSeek** released its open source R1 generative AI model in January 2025, raising data privacy, censorship, and national security concerns.

This article explores contrasting views in the evolving landscape of open source data and software use. It includes a discussion of potential risks related to cybersecurity and privacy, AI development, and intellectual property (IP) and licensing.

Regulators Caution Use of Open Source

U.S. regulators appear cautious regarding the use of open source data and software. For example, at a recent AI Roundtable hosted at the U.S. Securities and Exchange Commission (SEC), the benefits of open source were downplayed, while the risks were stressed. Additionally, the U.S. Department of Commerce and the Bureau of Industry and Security (BIS), have focused on potential national security implications of open source AI, particularly if controlled information is inadvertently released on open source.

- **Fear of Heightened Risk:** Like other software tools, open source AI raises regulatory concerns about cybersecurity, regulatory compliance, national security, and intellectual property. It's unclear how the Trump administration's ongoing **efforts to develop an "AI Action Plan"** will impact open source AI development and deployment.
- **Unclear Guidance:** In the recent SEC roundtable focused on AI, there was significant attention given to the risks associated with open source data and software and reservations from panelists about the merits of using open source. At the same time, innovation—especially in the AI marketplace—is being encouraged, and software engineers have for many years pointed to open source communities as valuable sources of innovation. Until guidance from agencies like the SEC is provided on the use of open source in AI tools, companies are on their own balancing the risks against the merits.
- **Potential Regulatory Action:** For example, SEC-regulated entities already have required **cybersecurity disclosures**. If the use of open source data and software is considered a cybersecurity risk, greater scrutiny may be placed on its use, and concomitant disclosures may need to occur. Also, the U.S. Department of Commerce and BIS have debated whether to add open source software to their export control product lists.

Combined, these efforts show an increased concern from U.S. regulators on the use of open source.

Advocacy for Open Source

Open source advocates include a broad coalition of intergovernmental organizations, such as the United Nations, as well as nongovernmental organizations (NGOs), academic institutions, and technologists. It also includes those engaged in digital transformation, and advocates for an expanded global commitment to open data and information sharing, especially in the AI space. Advocates focus on open source as a public good, for digital inclusion, and to advance tech sovereignty and global collaboration.

- **Public Good:** Advocates argue that open source information and software is a public good. Open source datasets—when properly documented and audited—enhance transparency, collaboration, and resilience against misinformation. In addition, open source software can advance innovation and help close digital divides.
- **Digital Inclusion:** Open source advocates also contend that open source datasets can help avoid the monopolization of foundational models and can help promote global digital inclusion. Many developers,

particularly in less tech-advanced economies, lack access to proprietary data sets, tools, or cloud infrastructure to compete in the global AI space. At the same time, many commercial AI systems are trained on English-language or Western-centric data systems, leading to underrepresentation and limited culturally contextual data.

- **Tech Sovereignty and Global Collaboration:** Some advocates are also drawn to open source AI as it facilitates global collaboration and peer review, especially among countries and startups trying to catch up to tech giants. Open source AI is also attractive to policymakers interested in “tech sovereignty,” as it enables tech development free from any one firm or government’s control. It allows stakeholders the opportunity to develop digital tools while reducing reliance on foreign tech firms.

Legal Risks

1. Cybersecurity Risks

Open source can also introduce significant cybersecurity risk, such as unknown contributors and unvetted code, potential vulnerabilities and a lack of security and technical support. These factors all contribute to whether, or how, cybersecurity laws and regulations apply to open source software. For example, while European Union policymakers created a new category for an organization to be classified as “open source software stewards” (**article 24**) in the CRA, there’s still uncertainty about who exactly is covered and what these (more limited) obligations will be given they’ll be developed in the next few years during implementation.

- **Unknown Contributors and Unvetted Codes:** Using open-source software essentially means incorporating someone else’s code into your own product or service. Anyone can contribute to open source projects, including threat actors. Therefore, open source code may be unvetted and include intentional backdoors, insecure details, or other behaviors that can introduce vulnerabilities.
- **Potential Vulnerabilities:** Open source software may not be regularly maintained and therefore may not be appropriately patched. Attackers frequently target known common vulnerabilities and exposures (CVEs) in widely used but outdated software. A vulnerability or malware in a single software can compromise the entire system.
- **Lack of Support:** Because many open-source data or software packages lack dedicated security teams or formal reviews, vulnerabilities may go undetected for long periods of time, exposing users to repeated exploits. Those who do maintain open source software do not always guarantee support, offering limited recourse in the event of a security incident.

2. AI Development and Use Risks

When using open-source data in AI development, there are layered concerns about “data poisoning,” lack of clear origin of records, and risk created by data integrity and data hygiene.

- **Data Poisoning:** In this newer type of cyber-attack, the threat actor may inject harmful or biased data into an open source dataset used to train AI models or decision-making systems, resulting in unpredictable or hidden behaviors.

- **Lack of Origin:** Open source datasets often lack clear records of origin, ownership, or methodology for how data was collected. This raises questions related to accuracy, legality, and compliance, and could result in unknowingly incorporating copyrighted, proprietary, or controlled information, as well as personally identifiable information (PII), personal health information (PHI), or geolocation data.
- **Data Integrity and Data Hygiene:** Open-source datasets may contain mislabeled, redundant, or outdated data, as well as data that is biased or discriminatory. AI models trained on low-quality data may result in discriminatory outputs, be less trustworthy, and be more susceptible to bias and manipulation.

3. Intellectual Property and Licensing Risks

Despite being “open,” using or contributing to open source datasets often carries IP risks, such as licensing compliance, attribution, and concern about maintaining potential trade secrets.

- **License Compliance:** Not all open datasets are a free-for-all. Some come with restrictive licensing requirements, which may be incompatible with certain use cases. Close attention should be given to the types of open source licensing that attaches to software or datasets that are made available online.
- **Derivative Works and Attribution:** Developers and companies using open data must assess whether outputs qualify as derivative works or require attribution under applicable licenses. Failure to confirm whether a license is required or attribution is required may result in financial liability risk and may impact business operations.
- **Trade Secrets:** Using publicly available datasets may put certain trade secrets at risk, particularly if data is not properly vetted. Therefore, incorporating open datasets can complicate IP enforcement.

Observations for Organizations Using Open Source Data

In light of these tensions, organizations should take a proactive and multidisciplinary approach to address the emerging risks related to the use of open source data.

- **Treat open-source datasets as part of the organizational risk landscape.** This means companies may consider conducting individual vetting on open source data and software, as well as implementing regular controls, including cybersecurity controls, to ensure regular security checks, patching, and data integrity verifications.
- **Establish an open source policy.** Open source policies may clarify uses, licensing, organizational roles and responsibilities, and legal processes to use open source data as well as to streamline secure software development. For example, it is important to ensure the quality of data used in certain datasets, as well as to assess whether an open source software option needs ongoing maintenance.
- **Establish policies for reviewing open source licensing.** Companies that use open source datasets in product development or AI training should be cautious to avoid inadvertent IP violations, which carry heavy litigation and potential regulatory risk.

Conclusion

While many see open source data and information as a cornerstone of inclusive, transparent digital innovation, regulators are increasingly cautious, emphasizing accountability, national security, and data integrity. Companies must navigate these competing pressures with care, drawing on legal, technical, and strategic perspectives.

Crowell & Moring LLP and Crowell Global Advisors are prepared to provide guidance and legal counsel around the use of open source data and information, particularly as it relates to cybersecurity, AI, and IP, and to help assess exposure under certain regulatory agencies.

For further information, please contact our team.

Contacts

Neda M. Shaheen

Associate

She/Her/Hers

Washington, D.C. D | +1.202.624.2642

nshaheen@crowell.com

Nigel Cory

Crowell Global Advisors Director

Washington, D.C. (CGA) D | +1.202.654.6753

ncory@crowellglobaladvisors.com

Jacob Canter

Counsel

He/Him/His

San Francisco D | +1.415.365.7210

jcanter@crowell.com