DO AS I SAY, NOT AS I CODE:

GITHUB'S COPILOT PROMPTS IP LITIGATION WITH INTERNATIONAL IMPLICATIONS

MARCO GERMANÒ*

I. INTRODUCTION

The rapid proliferation of Large Language Models (LLMs) in contemporary technological ecosystems has sparked significant legal debates, particularly regarding intellectual property (IP) rights.[1] LLMs are designed to detect patterns and relationships within vast datasets, often comprising a mix of publicly available and proprietary content. As these models process large volumes of textual data to generate outputs, concerns have emerged over their usage of proprietary materials—especially since LLMs typically do not distinguish between copyrighted and non-copyrighted sources, nor do they always seek authorization for their inclusion in training datasets. This increasing friction has been brought to light by high-profile cases, including lawsuits by The New

*Graduate Editor, *N.Y.U. Journal of International Law & Politics*.

1. *See, e.g.*, Simon Chesterman, *Good Models Borrow, Great Models Steal: Intellectual Property Rights and Generative AI*, POL'Y & SOC'Y (Feb. 12, 2024), https://doi.org/10.1093/polsoc/puae006 (analyzing copyright challenges in LLM training and human creativity); Michelle Lian, *Legal Frontiers: Navigating the Complex Landscape of Generative AI Regulation*, COLUM. U. L. REV. (May 9, 2024), https://www.culawreview.org/journal/legal-frontiers-navigating-the-complex-landscape-of-generative-ai-regulation (discussing copyright infringement and fair use in LLM training); Caitlyn Fernandes, *More AI, More Problems: The Legal Challenges Creatives Face in Uncharted Tech Territory*, N.Y.U. J. INTELL. PROP. & ENT. L. (Apr. 8, 2024), https://jipel.law.nyu.edu/more-ai-more-problems-the-legal-challenges-creatives-face-in-uncharted-tech-territory/ (addressing challenges faced by creatives due to generative AI's use of copyrighted data); Gil Appel, Juliana Neelbauer & David A. Schweidel, *Generative AI Has an Intellectual Property Problem*, HARV. BUS. REV. (Apr. 7, 2023), https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem (highlighting legal uncertainties in generative AI, including derivative works and fair use interpretations); Harry Surden, *ChatGPT, AI Large Language Models, and Law*, 92 FORDHAM L. REV. 1941 (2024), https://fordhamlawreview.org/wp-content/uploads/2024/03/Vol.-92_Surden-1941-1972.pdf (exploring the implications of LLMs in legal work).

York Times against OpenAI and by Getty Images against Stability AI.[2] While these cases are ongoing in U.S. courts, they are expected to shape artificial intelligence (AI) governance internationally, as the development and deployment of such technologies have been largely influenced by American regulatory standards.[3]

One particularly notable yet under-reported case is *Doe v. GitHub*, currently stayed in the Northern District of California after the court certified an order for interlocutory appeal on September 27, 2024.[4] This lawsuit involves OpenAI, GitHub, and its parent company Microsoft,[5] focusing on the use of open-source software (OSS) code to train LLMs, specifically GitHub's Copilot—a programming assistance tool currently powered by OpenAI's GPT-4 model and previously by Codex, a modified, fine-tuned version of GPT-3 additionally trained on gigabytes of publicly available source code.[6] Copilot allows developers to present coding problems in natural language or to input partial code in multiple programming languages, with the AI generating suggestions to autocomplete the entered code. Although praised for its potential to enhance programming productivity, open-source

2. Audrey Pope, *NYT v. OpenAI: The Times's About-Face*, HARV. L. REV. BLOG (Apr. 10, 2024), https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/; Hannah Ashbrook, *Getty Images Sues Stability AI: A Turning Point in AI Copyright*, FORDHAM INTELL. PROP. MEDIA & ENT. L.J. (Mar. 31, 2023), http://www.fordhamiplj.org/2023/03/31/getty-images-sues-stability-ai-a-turning-point-in-ai-copyright/.

3. *See* Kartik Hosanagar & Aneesh Ramesh, *Charting the Emerging Geography of AI*, HARV. BUS. REV. (Dec. 12, 2023), https://hbr.org/2023/12/charting-the-emerging-geography-of-ai (highlighting global competition in AI governance and the historical role of U.S. standards in shaping international technology deployment).

4. Doe 1 v. GitHub, Inc., No. 22-cv-06823, 2024 U.S. Dist. LEXIS 175951 (N.D. Cal. Sept. 27, 2024).

5. GitHub was founded in 2008 as a for-profit platform aimed at fostering collaboration in the open-source software development community. It quickly became an essential tool for developers due to its version control features, ability to host repositories, and collaborative coding capabilities. In 2018, Microsoft acquired GitHub for $7.5 billion, raising concerns within the open-source community due to Microsoft's previous opposition to open-source software. *See* Paul V. Weinstein, *Why Microsoft Is Willing to Pay So Much for GitHub*, HARV. BUS. REV. (June 6, 2018), https://hbr.org/2018/06/why-microsoft-is-willing-to-pay-so-much-for-github ("[Microsoft is] paying for the access it gets to the legions of developers who use GitHub's code repository products on a daily basis . . . so they can be guided into the Microsoft developer environment, where the real money is made.")

6. *The World's Most Widely Adopted AI Developer Tool*, GITHUB, https://github.com/features/copilot (last visited Oct. 4, 2024); *OpenAI Codex*, OPENAI, https://openai.com/index/openai-codex/ (last visited Oct. 4, 2024); Anthony Alford, *OpenAI Announces 12 Billion Parameter Code-Generation AI Codex*, INFOQ (Aug. 31, 2021), https://www.infoq.com/news/2021/08/openai-codex.

developers and communities have raised concerns about Copilot due to its tendency to reproduce material from public repositories without properly attributing authorship or adhering to terms and conditions of the original open-source licenses.[7]

At first glance, one might assume that open-source code does not carry the same legal risks as copyrighted material. However, open-source code is governed by licenses that set clear expectations for users and developers: they may generally use, modify, and redistribute the code for third-party use, but they must comply with the licensing terms, which often include requirements for attribution and copyright acknowledgment.[8] This attribution practice is a cornerstone of the open-source ethos, as illustrated by the fact that many licenses permit companies to use and build upon open-source code for commercial software, provided that attribution and copyright notices accompany any distributed versions of the licensed material.[9] Over the past few decades, this *licensing-as-governance* model has fostered a collaborative environment that promotes innovation and enables the swift correction of software malfunctions, even in the absence of the safeguards typically associated with closed, proprietary software. However, this model hinges on users adhering to the conditions set forth by the original licenses of the source code.

The plaintiffs in *Doe v. GitHub* argue that GitHub, Microsoft, and OpenAI failed to comply with open-source licensing conditions. This

---

7. *See, e.g.,* Craig Topham, *Publication of the FSF-funded white papers on questions around Copilot*, FREE SOFTWARE FOUNDATION (Feb. 24, 2022, 5:36 PM), https://www.fsf.org/news/publication-of-the-fsf-funded-white-papers-on-questions-around-copilot (raising concerns about copyright law, ownership of AI-generated code, and the legal impacts for GitHub developers using OSS licenses).

8. These licenses often aim to encourage innovation by ensuring that modifications and derivatives of the software remain open and accessible. However, it is important to differentiate OSS from free software, which, while sharing many principles with OSS, emphasizes users' rights to use, modify, and redistribute software without restrictions. While both OSS and free software promote openness, the former tends to focus on practical collaboration, whereas the latter is driven by ethical concerns about user autonomy and control. *See The Free Software Foundation (FSF)*, FREE SOFTWARE FOUNDATION, https://www.fsf.org/ (last visited Oct. 4, 2024) ("Free software means that the users have the freedom to run, edit, contribute to, and share the software. Thus, free software is a matter of liberty, not price."); *Defining Open Source AI*, OPEN SOURCE INITIATIVE, https://opensource.org/osd (Feb. 16, 2024) (proposing a 10-point definition of "open source").

9. *See* LAWRENCE ROSEN, OPEN SOURCE LICENSING: SOFTWARE FREEDOM AND INTELLECTUAL PROPERTY LAW (2005) ("Open source is now dominating many of the market conversations in the software industry. While software companies continue to release valuable and high-quality products under proprietary licenses, most are also embracing open source product development and distribution models as well as the software licenses that make those models possible.").

lawsuit marked the first U.S. class-action case to challenge both the training and output of AI systems that use open-source data, raising questions about how AI systems are trained using open-source code and the resulting implications for the open-source ecosystem. Notably, as the case has progressed, it has sparked new discursive disputes surrounding AI, copyright, and open-source governance. Its outcome has the potential to significantly influence the governance of AI, OSS, and IP law at a transnational level, with major implications for both the tech industry and the global open-source community.

This annotation will examine the case's central issues, recent developments, and broader implications for AI and copyright law. It will also explore how these connect to broader concerns about OSS legal governance and its role in the tech industry. The annotation is organized as follows: First, it outlines the facts and main arguments presented by both sides. Second, it analyzes the legal paths the Ninth Circuit may take and their potential consequences. Finally, it explores the broader debate on OSS transnational governance and its future. It concludes by suggesting potential avenues for further research.

II. THE CASE

In June 2022, following a year-long technical preview, GitHub officially released Copilot as a subscription-based service.[10] Promoted as an AI-powered coding assistant designed to suggest code and functions in real time, Copilot was marketed for being primarily trained on open-source code from public GitHub repositories.[11] While its release was highly anticipated, Copilot became entangled in legal controversies shortly after.

On November 3, 2022, attorney and open-source programmer Matthew Butterick, in collaboration with the San Francisco-based Joseph Saveri Law Firm—a firm specializing in generative AI-related litigation—filed a class-action lawsuit in the U.S. District Court for the Northern District of California.[12] Representing anonymous open-

---

10. Thomas Dohmke, *GitHub Copilot Is Generally Available to All Developers*, GITHUB BLOG, https://github.blog/news-insights/product-news/github-copilot-is-generally-available-to-all-developers (May 21, 2024).

11. Nat Friedman, *Introducing GitHub Copilot: Your AI Pair Programmer*, GITHUB BLOG, https://github.blog/news-insights/product-news/introducing-github-copilot-ai-pair-programmer/ (Feb. 23, 2022); Mark Chen et al., *Evaluating Large Language Models Trained on Code*, ARXIV, https://arxiv.org/abs/2107.03374 (Jul. 14, 2021).

12. *Greetings. This Is Matthew Butterick. I'm a Writer, Designer, Programmer, and Lawyer*, MATTHEW BUTTERICK, https://matthewbutterick.com (last visited Oct. 4, 2024); *GitHub and Copilot Intellectual Property Litigation*, SAVERI LAW FIRM,

source programmers, the lawsuit was directed at GitHub, Microsoft, and OpenAI. The plaintiffs raised 12 causes of action, including violations of the Digital Millennium Copyright Act (DMCA), breach of OSS licenses, tortious interference with contractual relationships, fraud, false designation of origin, unjust enrichment, and unfair competition. They also accused the defendants of privacy violations under the California Consumer Privacy Act (CCPA), negligence, and civil conspiracy, and sought declaratory relief for the alleged infractions.

Among the allegations, the core claim was that Copilot had been trained on billions of lines of open-source code without adhering to the licenses governing that code.[13] This alleged noncompliance formed the basis of claims under the DMCA, specifically § 1202(b), which prohibits the intentional removal or alteration of copyright management information (CMI)—the details that credit authors and specify license terms—without the copyright owner's consent.[14] By focusing on

---

https://www.saverilawfirm.com/our-cases/github-copilot-intellectual-property-litigation (last visited Oct. 4, 2024).

13. In *Doe v. GitHub*, the plaintiffs claimed that Copilot violated the terms of at least 11 OSS licenses, including prominent ones such as the Apache License, BSD License, GNU General Public License (GPL), MIT License, and Mozilla Public License (MPL). These licenses vary in their permissions and restrictions. The Apache License and MIT License are permissive, allowing broad usage with minimal obligations, mainly requiring attribution. The BSD License is similar but may include clauses preventing the use of the original author's name in promotional material. In contrast, the GNU GPL is a copyleft license, mandating that derivative works also be distributed under the GPL, ensuring that the software remains open-source. The MPL offers a balance, allowing integration with proprietary software but requiring that modifications to MPL-licensed code remain open-source. *See* Matthew Butterick, *Maybe You Don't Mind if GitHub Copilot Used Your Open-Source Code Without Asking*, GITHUB COPILOT INVESTIGATION, https://githubcopilotinvestigation.com (last visited Oct. 4, 2024) (explaining Copilot's potential violations of OSS licenses); *OSI Approved Licenses*, OPEN SOURCE INITIATIVE, https://opensource.org/licenses (last visited Oct. 4, 2024) (outlining the characteristics of OSS licenses, including those referenced in the investigation).

14. The DMCA, enacted in 1998, was designed to address copyright issues in the digital age and the rapidly evolving internet landscape. It provided strong protections against the circumvention of digital rights management (DRM) systems and introduced new legal tools for copyright holders to enforce their rights. A key provision relevant to these disputes is 17 U.S. Code § 1202(b), which addresses the "Integrity of Copyright Management Information" and includes the following prohibitions: (1) intentionally removing or altering any copyright management information, (2) distributing or importing for distribution the copyright management information, knowing that it has been removed or altered without the authority of the copyright owner or the law, or (3) distributing, importing for distribution, or publicly performing works, copies of works, or phonorecords, knowing that the copyright management information has been removed or altered without the authority of the copyright owner or the law, and

DMCA violations rather than traditional copyright infringement, the plaintiffs especially sought to address the harm caused by Copilot's failure to maintain proper attribution, which is a cornerstone of the open-source ecosystem.[15] They argued that Copilot violated this DMCA provision by autocompleting code governed by open-source licenses without proper attribution, depriving developers of rightful recognition. This lack of attribution, they emphasized, was not a mere oversight but a conscious disregard of legal obligations, with significant professional, economic, and reputational consequences within the open-source community, where recognition is vital for collaboration and career development.

In response, the defendants filed motions to dismiss the case on virtually all claims, specifically asserting that, regarding the DMCA claims, the plaintiffs had failed to identify instances where Copilot generated exact copies of their code. They argued that Copilot was not a "copy-paste" tool but rather an AI model that generated context-specific suggestions based on general programming patterns. Thus, Copilot did not engage in copyright infringement or the removal of CMI as alleged. The defendants also contended that the plaintiffs failed to state a valid claim, as GitHub's Terms of Service (TOS) provide broad rights to use, display, and reproduce code, thereby preempting the breach of license claims.

The District Court's initial ruling in May 2023 partially granted and partially denied the defendants' motions to dismiss but allowed the DMCA claims to proceed. The plaintiffs were then permitted to amend their complaint to provide more specific examples of Copilot's outputs that allegedly violated the DMCA. In their amended complaints, the plaintiffs doubled down on their DMCA claims, providing redacted examples of Copilot-generated outputs nearly identical to their original code. They argued that DMCA § 1202(b) should apply to collaborative and derivative works, not just identical copies, and maintained that the "fair use" doctrine was inapplicable in defending DMCA violations.[16]

---

knowing, or having reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement of any right under this title.

15. This strategy allowed the plaintiffs to address attribution issues without needing to prove traditional copyright infringement, which could have invited broader IP discussions, including those surrounding a "fair use" defense. Additionally, DMCA claims provide a clearer path to statutory damages.

16. The fair use doctrine is a legal principle under U.S. copyright law (17 U.S.C. § 107) that allows limited use of copyrighted material without requiring permission from the copyright holder. It is primarily intended to balance the interests of creators and the public, enabling transformative uses such as criticism, comment, news reporting, teaching, scholarship, or research. Courts consider four factors in determining whether

The plaintiffs emphasized that having to prove Copilot reproduced verbatim copies of their code would impose an unreasonable burden, particularly given the nature of generative AI models as systems that learn patterns from vast datasets to generate new content rather than storing or replicating data directly. Additionally, the plaintiffs expanded their allegations to include breach of contract claims related to GitHub's TOS. The defendants, however, continued to argue that the DMCA claims lacked merit, stressing that the plaintiffs had not demonstrated that Copilot produced exact copies of their code. They also contended that the plaintiffs failed to adequately identify the licensing provisions breached or specify instances in which CMI was removed. As such, the defendants argued, Copilot did not trigger reproduction without attribution.

On January 3, 2024, the court issued a second order in response to the defendants' motion to dismiss, granting parts of the motion while denying others. Regarding the DMCA dispute, the court concluded that for a claim under DMCA § 1202(b) to be valid, the works from which CMI is removed must be "identical" to the original, thereby dismissing plaintiffs' claims based on outputs in modified formats or functional equivalents. The plaintiffs' motion for reconsideration of the court's order was therefore denied. In response, the plaintiffs requested that the court certify its dismissal with prejudice of the DMCA claims for interlocutory appeal. They sought to challenge whether claims under § 1202(b)(1) or (b)(3) contain an "identicality" requirement, mandating that the AI output work must be identical to the original code to trigger a violation. The court granted this request, allowing the plaintiffs to appeal the DMCA ruling, and stayed proceedings pending a Ninth Circuit ruling on the matter.

III. THE CROSS-ROAD

The upcoming decision by the Ninth Circuit on whether claims under DMCA § 1202(b)(1) or (b)(3) must meet an "identicality" requirement carries significant implications for the AI industry, particularly in shaping standards for copyright compliance in models that use open-source data. A ruling that imposes strict requirements on how companies handle copyright metadata could influence the development, training, and deployment of AI systems across markets.

---

a use qualifies as fair: (1) the purpose and character of the use, including whether it is for commercial or non-profit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the whole work; and (4) the effect of the use on the potential market for or value of the copyrighted work. *U.S. Copyright Office Fair Use Index*, COPYRIGHT.GOV, https://www.copyright.gov/fair-use/ (last visited Oct. 4, 2024).

Importantly, as U.S. companies establish business models and digital infrastructure for AI technology internationally, this decision is likely to set de facto standards for AI governance worldwide, given the global reliance on U.S.-based AI technologies and legal frameworks.

While similar discussions are taking place regarding LLMs and copyright material in various U.S. and foreign courts, the discussion surrounding the DMCA is quite unique and puts the Ninth Circuit at a crossroads.[17] If the Ninth Circuit affirms that § 1202(b) claims require the removal or alteration of CMI from an "identical" copy of a work, plaintiffs in similar cases would face significant challenges in proving their claims. In *Doe v. GitHub*, the plaintiffs allege that Copilot reproduces portions of their licensed code without attribution, thus removing CMI. However, because Copilot generates modified, rather than verbatim copies, an identicality requirement would likely lead to the dismissal of the plaintiffs' DMCA claims.[18] Such a ruling could set a

---

17. Other jurisdictions are also confronting how AI models gather and use open-source and copyrighted data. In the European Union, the recently-approved "AI Act" includes specific rules on text and data mining. Wouter van Wengen & Radboud Ribbert, *EU AI Act's Opt-Out Trend May Limit Data Use for Training AI Models*, GREENBERG TRAURIG (July 3, 2024), https://www.gtlaw.com/en/insights/2024/7/eu-ai-acts-opt-out-trend-may-limit-data-use-for-training-ai-models. The United Kingdom is exploring legislative changes to balance research needs with creator rights. U.K. Intellectual Property Office, *The Government's Code of Practice on Copyright and AI*, GOV.UK (June 29, 2023), https://www.gov.uk/guidance/the-governments-code-of-practice-on-copyright-and-ai. Canada has introduced Bill C-27, which addresses privacy and data handling in a way that could also affect AI models trained on open-source and copyrighted materials. Innovation, Science and Economic Development Canada, *Bill C-27 Summary: Digital Charter Implementation Act, 2020*, GOVERNMENT OF CANADA, https://ised-isde.canada.ca/site/innovation-better-canada/en/canadas-digital-charter/bill-summary-digital-charter-implementation-act-2020 (Aug. 18, 2022). Meanwhile, several Asian countries are adjusting their legal frameworks to accommodate AI development; for example, China has instituted rules requiring generative AI services to comply with existing IP laws. Christopher Ferguson, Julie He & Dongwoo Kim, *China's New Rules for Generative AI*, FASKEN MARTINEAU DUMOULIN LLP (Aug. 29, 2023), https://www.fasken.com/en/knowledge/2023/08/chinas-new-rules-for-generative-ai. Japan has revised its copyright exceptions to clarify permissible data usage for training models. Aiko Yamada & Yuki Sako, *Japanese Government Identified Issues Related to AI and Copyrights*, NAT'L L. REV. (Sept. 26, 2023), https://www.natlawreview.com/article/japanese-government-identified-issues-related-ai-and-copyrights. Although these discussions share common themes, the DMCA-specific dimension of Doe v. GitHub stands out for its focus on whether an "identicality" requirement applies under § 1202(b), potentially redefining how open-source licensing and attribution obligations interact with AI-generated outputs.

18. Should the Ninth Circuit uphold the "identicality" requirement, and consequently dismiss the DMCA claims, the plaintiffs would still have avenues to pursue their remaining claims. They could focus on breach of contract claims related to the

precedent that narrows the DMCA's application, enabling developers to argue that AI-generated outputs derived from copyrighted works are legally distinct. This would encourage further reliance on publicly available datasets for AI training at the expense of appropriate attribution to the original developers. It would also impact the professional dynamics of the open-source ecosystem by encouraging AI use for productivity gains, though it would likely necessitate a reorganization of licensing practices.

Conversely, if the Ninth Circuit rejects the "identicality" requirement, the plaintiffs would only need to demonstrate that CMI was removed or altered from a derivative or modified version of their code. This broader interpretation of § 1202(b) would expand DMCA protections, allowing plaintiffs to assert their rights over AI-generated outputs that lack proper attribution. Such a decision could lead to increased scrutiny of AI training processes and outputs, imposing stricter obligations on developers to ensure compliance with IP protections, even for altered or derivative works. This outcome would significantly reshape the global development and deployment of AI tools, as developers worldwide may face increased legal risks when using open-source code in AI models, prompting a shift toward alternative training or proprietary datasets to mitigate litigation risks. The rationale adopted in this case could also influence other legal discussions on LLM training practices and the boundaries of using open-source data.

IV. THE BROADER DEBATE

The ongoing *Doe v. GitHub* case taps into two emerging conversations on a broader scale. One, perhaps more obviously, is the regulation of AI systems, including debates on openness, transparency, and their relationship with copyright law and safeguards. The other, less noticed, concerns the global governance of OSS ecosystems.

Over the past few decades, OSS has flourished, providing a reliable and cost-effective solution for developers, companies, and governments, and becoming ingrained in global digital infrastructure.[19] This

---

failure to include attribution and license terms as required by OSS licenses, as well as breach of GitHub's TOS concerning the sale of materials without proper licensing compliance. Although the court found certain state law claims to be preempted by the Copyright Act, the plaintiffs also may still seek restitution or remedies through breach of contract theories, provided they meet the legal standards for such claims. These alternative strategies could result in damages or injunctive relief for the plaintiffs.

19. Studies estimate that the cost to develop widely used OSS—the supply-side value—is approximately $4.15 billion, while the value derived by users—the demand-side value—reaches around $8.8 trillion. Without OSS, firms might have to spend up

ecosystem has developed through a unique governance model, primarily based on licenses and operating largely outside conventional legal frameworks.[20] Recently, however, this model has been increasingly tested, with OSS supply chain attacks rising by 742 percent over the last three years, marked by the injection of more than 245,000 instances of malicious code into open-source projects—some of which have targeted critical infrastructure in the United States.[21] These attacks often compromise security by stealing data, introducing vulnerabilities, or allowing unauthorized system access.[22] This stark rise in threats has raised questions about whether the current OSS governance structure—largely reliant on licenses and community enforcement—is legally and practically sufficient. The very openness that makes OSS successful also makes it vulnerable to security risks and legal ambiguities.

---

to 3.5 times more on software, significantly impacting their profitability and capacity for innovation. Manuel Hoffmann, Frank Nagle & Yanuo Zhou, *The Value of Open Source Software* (Harv. Bus. Sch. Working Paper No. 24-038, 2024), https://www.hbs.edu/faculty/Pages/item.aspx?num=65230. Indeed, it is estimated that at least 90% of companies use open-source software, and 97% of commercial codebases incorporate open-source components. Annamaria Conti, Vansh Gupta, Jorge Guzman & Maria P. Roche, *Incentivizing Innovation in Open Source: Evidence from the GitHub Sponsors Program* (Nat'l Bureau of Econ. Rsch., Working Paper No. 31668, 2023), https://www.nber.org/be/20241/open-source-software-creators-its-not-just-about-money.

20. David McGowan, *Legal Implications of Open-Source Software*, SSRN (Dec. 2, 2000), https://ssrn.com/abstract=243237.

21. By *supply chain*, I refer to the sequence of actors and processes involved in the OSS lifecycle, starting with the project's initial development by *developers*, followed by code modifications made by *contributors*, oversight from *project managers*, distribution through public *repositories*, and finally, its deployment in applications by individual, corporate, and governmental *users* worldwide, who may further provide products to *consumers*. Notable incidents include the 2018 Event-Stream attack, where a popular Node.js package was compromised with malicious code; the 2021 Log4Shell critical security vulnerability, where the widely-used Apache Log4j library exposed millions of systems to remote code execution attacks; and the 2024 XZ Utils attack, where a malicious backdoor was introduced into the Linux utility xz within the liblzma library. *See* NADIA EGHBAL, WORKING IN PUBLIC: THE MAKING AND MAINTENANCE OF OPEN SOURCE SOFTWARE (2020); *A History of Software Supply Chain Attacks: July 2017–Present*, SONATYPE, https://www.sonatype.com/resources/vulnerability-timeline (last visited Oct. 4, 2024) (providing a timeline of attacks since July 2017).

22. Paolo Mainardi, *The Rising Threat of Software Supply Chain Attacks: Managing Dependencies of Open Source Projects*, LINUX FOUND. (Aug. 15, 2023), https://linuxfoundation.eu/newsroom/the-rising-threat-of-software-supply-chain-attacks-managing-dependencies-of-open-source-projects.

These incidents have prompted actors worldwide to revisit OSS governance models and security practices.[23] Proposals range from scrutinizing the use of OSS in critical systems to broadening governmental oversight of software providers.[24] At the same time, public and private stakeholders increasingly recognize the need for globally coordinated technical safeguards, such as stricter contributor authentication, tighter access controls, regular security audits, and greater financial support for maintaining essential OSS projects.[25] Indeed, because open-source

23. *See, e.g.*, Exec. Order No. 14,028, 86 Fed. Reg. 26,633 (May 12, 2021), https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/ (calling for new regulations to ensure and attest "to the integrity and provenance of open source software used within any portion of a product"); European Union Agency for Cybersecurity, *Threat Landscape for Supply Chain Attacks* (July 29, 2021), https://www.enisa.europa.eu/publications/threat-landscape-for-supply-chain-attacks (recommending actions to mitigate supply chain threats arising from OSS); *Understanding and Responding to the SolarWinds Supply Chain Attack: The Federal Perspective: Hearing Before the S. Comm. on Homeland Sec. & Governmental Affs.*, 117th Cong. (2021), https://www.hsgac.senate.gov/hearings/understanding-and-responding-to-the-solarwinds-supply-chain-attack-the-federal-perspective/ (highlighting the need for enhanced security practices along the OSS transnational supply chain).

24. *See, e.g.*, Christian Vasquez, *White House to Study Open Source Software in Critical Infrastructure*, CYBERSCOOP (Aug. 9, 2024), https://cyberscoop.com/open-source-critical-infrastructure-def-con (discussing the Biden administration's creation of an office within the Department of Homeland Security to analyze the security of OSS in critical infrastructure, with support from the Department of Energy's national laboratories); Cyber Safety Review Board, *Review of the December 2021 Log4j Event* (July 11, 2022), https://www.cisa.gov/sites/default/files/publications/CSRB-Report-on-Log4-July-11-2022_508.pdf (analyzing expanding the Cybersecurity and Infrastructure Security Agency's (CISA) role in cyber risk communication, strengthening regulatory enforcement of security guidance, enhancing Software Bill of Materials (SBOM) tools, creating a government-led group to track software vulnerabilities, exploring software transparency requirements for federal vendors, among other measures); Wally Adeyemo, *How to Stop Cyberattacks on the U.S. Financial System*, BLOOMBERG (Jan. 17, 2025), https://www.bloomberg.com/opinion/articles/2025-01-17/how-to-stop-cyberattacks-on-the-us-financial-system (arguing that cyberattacks pose the greatest risk to the U.S. financial system, advocating for expanded Treasury Department authority to regulate third-party service providers, improve cyber intelligence sharing, and enhance coordination among financial regulators).

25. *See, e.g.*, Press Release, Open Source Software Security Foundation, The Linux Foundation and Open Source Software Security Foundation (OpenSSF) Gather Industry and Government Leaders for Open Source Software Security Summit II (May 12, 2022), https://openssf.org/press-release/2022/05/12/the-linux-foundation-and-open-source-software-security-foundation-openssf-gather-industry-and-government-leaders-for-open-source-software-security-summit-ii/ (outlining a 10-point strategy to enhance open source security, focusing on secure software development education, risk assessment, digital signatures, memory safety, incident response, vulnerability scanning, code audits, data sharing, widespread adoption of SBOMs, and improved

projects and communities operate across national borders, any governance reform must account for these transnational considerations, emphasizing that OSS infrastructure is inherently international rather than limited to local jurisdictions.

The *Doe v. GitHub* case adds to these broader governance issues and resonates with global discussions about AI-powered tools' reliance on publicly accessible code. The case's ruling on how the DMCA applies to AI-generated or modified code will likely set de facto standards for open-source licensing and compliance, shaping how developers, companies, and governments address OSS use in AI systems. The decision will also become part of a critical moment in which regulators are addressing the implications of AI systems trained on vast volumes of open-source code. These systems introduce new vulnerabilities into the transnational software supply chain, with flaws in the training data potentially affecting users far beyond national boundaries. This raises pressing questions about liability if an AI copilot's suggestions lead to malicious code: should responsibility lie with the copilot itself, GitHub as the host, the user deploying the code, or the original developer whose compromised code was used in training? These issues of responsibility and governance gaps within the OSS ecosystem are central to its transnational deployment and require future exploration.

V. CONCLUSION

The outcome of the *Doe v. GitHub* case could set a major precedent for AI-generated content and copyright law, particularly in the tech sector, where open-source licenses are widely used. The resolution is likely to shape how AI tools use OSS without violating licensing agreements, potentially leading to stricter IP compliance rules or reaffirming the flexibility of using open-source material with fewer constraints. Beyond AI and copyright law, the case underscores governance challenges within the OSS ecosystem. As OSS has become essential to global digital infrastructure, its decentralized nature has exposed legal and security gaps, particularly in enforcing licenses and safeguarding against vulnerabilities. The rise of AI models trained on

---

supply chain security); NAT'L INST. OF STANDARDS & TECH., FRAMEWORK FOR IMPROVING CRITICAL INFRASTRUCTURE CYBERSECURITY, VERSION 1.1 (2018), https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf (last visited Oct. 4, 2024) (introducing voluntary standards on authentication and identity, self-assessing cybersecurity risk, vulnerability disclosure, and supply chain risk management); NADIA EGHBAL, WORKING IN PUBLIC: THE MAKING AND MAINTENANCE OF OPEN SOURCE SOFTWARE (2020) (emphasizing the critical need for increased financial support to maintain essential OSS projects, highlighting their role as vital infrastructure that requires ongoing maintenance).

open-source datasets adds complexity to this landscape, raising challenging questions about how OSS governance can evolve to strengthen accountability, security, and openness in a rapidly evolving transnational environment.