Copyright Issues in Generative AI for Software: *Doe v. Github Inc. et al.*

Jeffrey W. Gluck*

Abstract: *Generative artificial intelligence (AI) allows users to generate content based on providing prompts to the generative AI. There are generative AI programs now capable of generating different types of content, including software code. The ongoing* Doe v. Github Inc. et al. *litigation addresses copyright-related issues inherent in the Copilot generative AI that allows users to enter prompts to generate software code. This case addresses many of the issues involved in the training and use of generative AI for generating software code.*

Introduction

The latest trend in the world of artificial intelligence (AI) is generative AI (primarily generative neural networks (GNNs)) for creating works. Already, GNNs exist to generate text (e.g., ChatGPT), images (e.g., DALL-E), and software (e.g., Copilot). All GNNs require training data for learning how to perform their tasks. Once trained, upon user prompts, these GNNs generate works in line with the user prompts. However, it is also notable that lawsuits have been filed based on such GNNs, where many of the issues are rooted in copyright law. In the specific case of software there is an ongoing case in the U.S. District Court for the Northern District of California,[1] in which many of the accusations are either directly or indirectly rooted in copyright law.

The Case of *Doe v. Github*

Doe v. Github is a class action lawsuit filed by a number of software developers (on behalf of the class of similarly situated software

Journal of Emerging Issues in Litigation / Winter 2024, Vol. 4, No. 1, pp. 29–35.

ISSN 2835-5040 (print) /ISSN 2835-5059 (online).

developers) who posted computer programs they developed, in the form of source code, on the open-source software depository run by Github Inc. ("open-source software" is source code (i.e., the instructions as written by a programmer) that is subject to permission to be copied and reused, generally under the conditions of an accompanying open-source license). The lead plaintiffs used various open-source licenses that required, at the very least, attribution of the code to the developer, as well as retaining the licensing terms with the code (in some cases, this might have included a link to the specifics of the license or simply the name of the license under which the source code was to be distributed).

Copilot is a GNN-based add-on for programming platforms. Copilot was developed jointly by Github and OpenAI, while OpenAI developed Codex, the underlying GNN-based AI model incorporated into Copilot. Microsoft owns Github and has made major investments in OpenAI; additionally, Copilot runs on the Microsoft Azure cloud computing platform. Github Inc., Microsoft Corporation, and various related OpenAI entities are the defendants in *Github*. The central themes underlying the lawsuit are that Copilot was trained using source code obtained from Github and subject to open-source licenses and that outputs from Copilot when prompted to generate code for the same purposes as the original code used for training have been the same or substantially similar to the original code. It is a further theme that the generated code lacked attribution of authorship or other licensing-/copyright-related information.

Currently, following an initial decision based on an initial complaint filed on November 3, 2022, the court issued a decision,[2] ruling on the defendants' motions to dismiss the complaint. The court granted these in part and denied them in part. Most of the dismissals granted plaintiffs the opportunity to file amended claims for relief, and therefore, the plaintiffs filed an amended complaint on June 8, 2023. It is noted that none of the pending claims for relief of the complaint is for copyright infringement, *per se*; however, copyright law is implicated in a number of the claims for relief, and it is instructive to review these for how copyright issues are implicated.

The first claim for relief in the amended complaint alleges violations of the Digital Millennium Copyright Act (DMCA), specifically 17 U.S.C. §§ 1202(b)(1) and 1202(b)(3). The DMCA was passed in 1998 and amended copyright law to address measures to prevent copyright infringement. Section 1202 specifically addresses maintaining the integrity of copyright management information (CMI), which is defined in the 17 U.S.C. § 1202(c) to include copyright notices, information about the author and/or copyright owner, and other similar information, which may depend on the nature of the work and/or the medium in which the work is conveyed.[3] Section 1202(b)(1) specifically prohibits intentionally removing or altering any CMI, and § 1202(b)(3) prohibits distribution and related acts, knowing that CMI has been removed or altered without authority of the copyright owner.[4] This is part of the body of U.S. copyright law, and the underlying notion here is that the work must be subject to copyright protection. Copyright protection adheres when a work is "fixed in a tangible medium of expression."[5] Therefore, as soon as the programmer writes a program on paper, stores it in a storage device, or the like, it is subject to copyright protection. In the case of open-source software, the software developer chooses among many available open-source licenses (or may even make up their own), having a wide range of terms. The open-source license is generally set forth along with the source code and most often includes the name of the author and/or copyright holder, if the copyright holder differs from the author. In this way, copyright law is central to the DMCA, that is, in that the DMCA addresses publication and enforcement of licenses based on the work being subject to copyright; stated another way, if there is no copyright, there is no basis for enforcement of the DMCA.

In *Github*, the plaintiffs assert that Copilot was trained using open-source software deposited to Github and that Copilot outputs source code that is identical to or similar to source code that they authored, which they deposited to Github; which was subject to an open-source license; and which does not maintain the open-source license information included with the original source code (on a related note, the plaintiffs also point out that Github offers a selection of open-source licenses that authors may append to their source code, including the most popular open-source licenses, all

of which include copyright notices). They support this with specific examples of code that they deposited on Github and code that was generated by entering prompts to generate source code serving the same purpose as the code that they had written. The plaintiffs point out that in each case the output of Copilot was a "verbatim," "almost verbatim," or "essentially verbatim" copy of the original source code or was a "modified form" of the original source code. In no case was the open-source licensing information reproduced with the code generated by Copilot. Additionally, the plaintiffs provide proof that earlier versions of Copilot did output open-source license information and that later versions of Copilot were trained to omit this information. The latter was provided by plaintiffs to address the defendants' knowledge and intentions regarding CMI, where the CMI constitutes the open-source license information. The allegations are that Github, Microsoft, and OpenAI all were aware of the above and, by offering Copilot in its present form, participated in acts of intentionally removing CMI from the source code and in acts of distributing source code subject to CMI in which the CMI was removed without authorization.

There are two types of open-source licenses, permissive and restrictive. A permissive open-source license allows one to copy and use the associated source code, generally with the only condition being that the license/copyright information, including attribution, continue to be appended to the source code. A restrictive (or "viral") open-source license not only requires the same conditions as a permissive license but also requires that any source code that incorporates the open-source code (or in some cases, makes other use of the open-source code) must be subject to the same license as the open-source code (one of the implications of this is that the software that incorporates open-source software subject to such a license must, in turn, be offered as open-source software).

Accordingly, in a second claim for relief, the plaintiffs assert violation of the open-source licenses that were attached to their software as a common-law breach of contract. Among the details of the claim are that both permissive and restrictive open-source licensing terms were breached for various open-source software of the various plaintiffs. The basis of this claim is that one who copies open-source software accepts the terms of the license and that, by

not adhering to the terms of the licenses, the defendants breached a license (i.e., a contract). It is again noted that without copyright, the licenses/contracts would be null and void, as plaintiffs would have had nothing to offer as part of the licenses/contracts.

While there were six other claims for relief asserted in the amended complaint, the above two were those most strongly based in copyright law.

As noted above, no direct violations of copyright were asserted among the claims for relief, either in the original complaint or in the amended complaint; all copyright-related claims were based on either DMCA provisions, having to do with rights management, or breach of contract/license. However, other pending lawsuits regarding other types of copyrighted works have made copyright infringement claims in connection with GNNs. For example, comedian, actress, and writer Sarah Silverman, along with other authors, have sued OpenAI for direct and vicarious copyright infringement under 17 U.S.C. § 106 in connection with their GNNs, evidenced based on their ability to output a detailed synopsis of every part of a book, asserting that these GNNs must have been trained on unauthorized copies of their books.[6] Direct infringement is alleged in that OpenAI's GNNs must be retaining "expressive information extracted from Plaintiff's works (and others) and retained inside them," thus making the GNNs themselves infringing derivative works.[7] The vicarious infringement claim is based on outputs of the GNN being infringing derivative works.[8]

It is curious that the same attorney representing Sarah Silverman and her co-plaintiffs is also representing the "Does" in *Github*, and yet the same or similar claims were not made in the *Github* complaint. While it is not clear, perhaps the reasons may relate to the different natures of the copyrighted works in *Github* and in *Silverman*. In particular, U.S. copyright law includes specific limitations on exclusive rights in computer programs.[9] However, it would appear that these limitations do not apply under the circumstances of *Github*. That is, it may be argued, on the same bases as in *Silverman*, that Copilot may also be viewed as a derivative work and that it outputs unauthorized copies and/or derivative works.

Or perhaps there is a far deeper reason why copyright infringement is not raised in *Github*. Section 102 of the copyright statute

states, "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work."[10] Software inherently embodies a procedure, process, and method of operation (of a computing device). Herein lies an inconsistency within U.S. copyright law. U.S. copyright law was amended in 1980 to specifically add § 117 (cited above) and to add language to apply copyright law to computer programs. However, § 102(b) was not amended to create any exceptions, nor did any of the amendments specifically state an exception to § 102(b). This leads to a quandary for a court: should § 102(b) be read strictly to exclude computer programs, or should the amendments made in 1980 be implicitly understood as limiting § 102(b)? A case in point is the Supreme Court decision in *Google v. Oracle*.[11]

In *Google*, the question was whether Google had infringed Oracle's copyright in software relating to the Java application programming interface (API).[12] The majority opinion avoided the question of copyright eligibility for computer programs and ruled in favor of Google based on the reasoning that if copyright attached to computer programs, the particular circumstances of the case dictated in favor of fair use by Google, applying § 107 of the copyright statute.[13] The dissenting opinion cited every instance in which language relating to computer programs was mentioned in the copyright statute and concluded that copyright law did apply to computer programs and, in particular, to the Java API.[14] Perhaps the majority skirted the issue because of a disagreement as to the applicability of copyright law to computer programs. In any event, given the result in *Google*, one might be dissuaded from attempting to pursue a "pure copyright" claim for computer programs.

Conclusion

The current status of *Github* is that the amended complaint mentioned above has been filed, and as of the date of the writing of this article, the defendants have not yet filed responses. It will be interesting to see how this case evolves and what the outcome

is. *Github* is a case that may have far-reaching implications for AI-generated works in the future.

Notes

* Jeffrey W. Gluck (jgluck@panitchlaw.com) is a Partner at Panitch Schwarze Belisario & Nadel LLP. His legal practice spans a broad range of areas, including patent application and prosecution, IP counseling, providing legal opinions, serving on litigation teams, and appellate litigation.

1. *Doe v. Github, Inc. et al.*, Case Nos. 4:22-cv-06823 and 4:22-cv-07074 (N.D. Cal.).
2. *Doe v. Github, Inc.*, 2023 U.S. Dist. LEXIS 86983 (N.D. Cal. 2023).
3. 17 U.S.C. §§ 1202(b) and 1202(c).
4. 17 U.S.C. §§ 1202(b)(1) and 1202(b)(3).
5. 17 U.S.C. § 102(a).
6. *Silverman v. OpenAI, Inc.*, 3.23-cv-03416 (N.D. Cal.).
7. Complaint at 12, *Doe v. Github, Inc.*, 2023 U.S. Dist. LEXIS 86983 (N.D. Cal. 2023) (Case No. 22-cv-06823).
8. *Id.*
9. 37 U.S.C. § 117.
10. 17 U.S.C. § 102(b).
11. *Google LLC v. Oracle America, Inc.*, 593 U.S. ___, 141 S. Ct. 1163 (2021).
12. *Id.*
13. *Id.*
14. *Id.*