Compliance Challenges with AI-based Code Generation Tools

Samriddhi

Northeastern University, Khoury College of Computer Sciences, 360 Huntington Ave, Boston, MA, USA 02115 lnu.samr@northeastern.edu

Abstract

The rapid development of AI-based code generation tools, such as Amazon CodeWhisperer and Microsoft Copilot, presents both opportunities and challenges for software development companies aiming to maintain compliance. This paper explores these issues by addressing several key questions. Firstly, it examines the data used to train the AI algorithms that support these code generation tools, focusing on the precautions taken by vendors to mitigate the risk of biased or malicious training data, which could result in non-compliant or vulnerable code. Secondly, the paper investigates the implications of AI-generated code for intellectual property (IP), including ownership rights and licensing issues. It emphasizes the need for businesses to ensure they have the correct licenses in place before using and distributing AI-generated products. Lastly, to ensure the secure and reliable integration of AI code generation tools into corporate operations, the paper addresses the necessary vendor due diligence. This involves evaluating suppliers' security protocols, accountability procedures, and transparency policies to prevent compliance violations. This study aims to guide stakeholders in effectively balancing the benefits of AI-driven innovation with the imperative of regulatory compliance in software development.

1. Introduction

The use of Artificial Intelligence (AI) to assist in the creation of computer programs is increasing day by day. While this practice can aid in software development and improve coding proficiency, it does not guarantee security and compliance. Utilizing automation for code

generation is a double-edged sword. In fact, recent studies indicate that some AI models generate software with vulnerabilities. (1)

Another significant issue is the rights to intellectual property (IP). According to Dehouche (2021), the question of who owns the code produced by AI tools is complex and multifaceted, involving copyright law implications as well as the potential for unintentional code plagiarism. To avoid intellectual property violations and ensure compliance with existing legal frameworks, organizations must carefully navigate these legal complexities. (2)

Furthermore, thorough vendor due diligence is necessary before deploying AI-based code generation solutions. To ensure secure and reliable commercial operations, organizations must assess the security, dependability, and ethical implications of these tools, especially considering challenges like data privacy and confidentiality issues. This is particularly crucial in regulated industries where adherence to strict moral and legal requirements is mandatory.

The purpose of this paper is to thoroughly examine these compliance issues and offer suggestions and insights to enterprises looking to effectively utilize AI-powered code generation technologies. This study contributes to a better understanding of the compliance landscape by exploring the intersection of technology, law, and ethics, and it provides recommendations for overcoming the challenges associated with developing AI-driven software.

2. Related Work

Several studies and reports have explored the significant risks involved in the domain of AI-generated tools, particularly in the context of software development and cybersecurity.

According to a report by the U.S. Department of the Treasury, data poisoning, data leakage, and data integrity attacks can occur at any stage of the AI development and supply chain, which can have adverse effects on financial institutions. The study suggests that AI systems are more vulnerable to these concerns than traditional software systems due to their dependency on the data used to train and test them. When AI systems ingest data during training or testing, this

can directly influence the production processing of the AI system. Source data, training datasets, testing datasets, pre-trained AI models, Large Language Models (LLMs), prompts, and vector stores can all be subject to data attacks. Therefore, emphasizing the importance of securing data throughout the development and production cycle. (3)

In a lawsuit filed in late 2022, Andersen v. Stability AI et al., three artists initiated a class action against several generative AI platforms. They claimed that these platforms used their original works without permission to train AI models in their artistic styles, enabling users to create works that may not be sufficiently different from their copyrighted pieces, thus constituting unauthorized derivative works. This improper use of the original work violated the copyright and trademark of the individuals' intellectual property. (4)

A study conducted by the Ponemon Institute in 2019 found that 59% of companies experienced a data breach caused by a vendor within the past year, underscoring the critical need for comprehensive vendor assessments. The study emphasizes that inadequate vendor due diligence can expose organizations to significant financial, compliance, and reputational risks. It highlights the necessity for businesses to conduct thorough evaluations of their vendors' cybersecurity measures, business continuity plans, and compliance with regulations to mitigate these risks effectively. The findings suggest that many organizations lack awareness of the extent of their exposure to third-party risks, with nearly 25% unaware of whether they have been impacted by a vendor's data breach. (5)

3. Analysis

AI presents ethical and legal challenges, particularly when generating code for handling sensitive data or for safety-critical systems. Numerous obstacles must be overcome, and rigorous validation and verification procedures are needed to ensure that the generated code complies with security, privacy, and ethical standards. To ensure responsible and safe use, ethical issues, biases, and potential hazards in AI-generated code must be addressed. AI models that generate code might

handle private or confidential data, making it critical to protect user privacy and ensure secure code production. Development teams need to implement strong security measures to safeguard sensitive data and prevent the malicious use of AI-generated code.

Compliance with data protection regulations, such as GDPR, is a significant challenge for AI systems that generate code involving personal data. To prevent AI tools from violating privacy laws, robust data governance and security measures are required. AI-generated code must be secure to prevent malicious use or data breaches. Implementing strong security protocols and conducting regular audits can help safeguard sensitive data and prevent unauthorized access. (6)

AI models often generate code without a full understanding of the specific security requirements of a given application or environment. This can lead to code that does not adhere to best practices for security compliance. Vulnerabilities like SQL injection, cross-site scripting (XSS), and buffer overflows can be unintentionally introduced by AI-generated code, especially if the training data contains insecure coding patterns. Additionally, sensitive data and intellectual property may be at risk from AI techniques. Any intellectual property or client information entered into the tool may also be saved or viewed by other service providers. Furthermore, any sensitive information inputs may wind up in outputs for other users because data submitted into generative AI tool prompts might also become part of its training set. Although this might seem like a low-risk situation, 11% of the data that employees input into ChatGPT is confidential, according to a recent Cyberhaven research. (7)

Not all code generated by AI tools is security compliant. Unlike human developers who can apply security frameworks and guidelines during coding, AI tools do not inherently follow security standards like OWASP (Open Web Application Security Project) or specific regulatory requirements like PCI-DSS (Payment Card Industry Data Security Standard). This means that code produced by AI might not comply with industry-specific security requirements unless explicitly guided by human oversight.

4. Recommendations

Organizations implementing AI-based code generation tools must put in place a complete set of strategies to reduce the related risks, given the compliance concerns detailed in this paper.

Some government and regulatory bodies are proactively proposing or enacting legislation to regulate the development and use of AI. The European Union's Artificial Intelligence Act (8) is currently in a provisional agreement stage and is expected to be enacted by 2025. In the United States, President Biden's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence has established the U.S. Artificial Intelligence Safety Institute (USAISI) (9). This institute will spearhead efforts to create standards for the safety, security, and testing of advanced AI models. In the absence of comprehensive federal legislation, it is anticipated that there will be industry-specific actions from agencies across sectors such as healthcare, financial services, housing, workforce, and child safety, along with further executive orders. The growing use of AI tools and their impact on organizational risk have led to the creation of several new AI security frameworks. For instance, ISO 42001 (10) provides a broad approach to information security applicable to AI systems, while the NIST AI RMF (11) offers detailed guidelines for managing risks specifically associated with AI technologies. By evaluating and adopting these frameworks, organizations can better navigate the complex landscape of AI risks and implement AI solutions in a secure and responsible manner.

When evaluating vendors that provide AI systems or AI-dependent products and services, institutions should consider broadening their standard third-party due diligence and monitoring to address AI-specific concerns. Beyond typical third-party risk assessments, institutions should inquire about AI technology integration, data privacy, data retention policies, AI model validation, and AI model maintenance. For example, initially, Slack did not explicitly ask for customer consent before using their data to train AI models, which led to backlash from users. The company's privacy policy indicated that it could use customer data by default for training its AI

models unless users opted out. This default opt-in setting was criticized, as it meant that user messages, files, and other shared content could be used for AI training without explicit permission. After facing criticism, Slack updated its privacy policies to clarify that it does not use customer data to train third-party AI models. This example highlights the importance of transparency and consent in the use of AI tools for data processing.

5. Conclusion

AI holds significant potential and can greatly enhance efficiency for developers, but it also introduces complex legal challenges. However, there are strategies to manage these legal risks. It is wise for companies to create or update their AI usage and open source policies to ensure responsible use by employees and contractors.

AI code generation tools like Copilot and CodeWhisperer should complement skilled human developers, rather than serve as standalone solutions. While these tools can greatly enhance productivity by generating code snippets, suggesting improvements, and automating repetitive tasks, human oversight is essential to ensure compliance with security and ethical standards.

Human developers bring context, critical thinking, and a deep understanding of specific project requirements that AI currently lacks. For instance, AI might generate code that technically works but doesn't align with security best practices or organizational guidelines. Therefore, developers must validate and refine AI-generated code to ensure it meets compliance requirements, particularly regarding data privacy, security, and intellectual property rights.

Moreover, compliance considerations are crucial when deploying AI code generation tools, particularly concerning how the AI models are trained and where data is stored. The data used to train these AI models can include sensitive or proprietary information, raising concerns about data privacy and intellectual property. It's important to verify that the training data used by these tools has been obtained and processed ethically and legally, with proper consent where necessary.

Additionally, organizations should be aware of where this data is stored and how it is protected to prevent unauthorized access or data breaches.

In conclusion, while AI-driven tools offer significant benefits, they must be used with careful consideration of compliance and security, ensuring they enhance rather than compromise the integrity of software development.

6. References

(1) A systematic literature review on the impact of AI models on the security of code generation
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11128619/

(2) Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3)
https://www.int-res.com/articles/esep2021/21/e021p017.pdf

(3) Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector U.S. Department of the Treasury March 2024
https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf

(4) Artists and Illustrators Are Suing Three A.I. Art Generators for Scraping and 'Collaging' Their Work Without Consent
https://news.artnet.com/art-world/class-action-lawsuit-ai-generators-deviantart-midjourney-stable-diffusion-2246770

(5) 2019 Global State of Cybersecurity in Small and Medium-Sized Businesses
https://www.cisco.com/c/dam/en/us/products/collateral/security/ponemon-report-smb.pdf

(6) AI Chatbots and Challenges of HIPAA Compliance for AI Developers and Vendors
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10937180/

(7) Risk and Compliance in the Age of AI: Challenges and Opportunities

https://secureframe.com/blog/ai-in-risk-and-compliance

(8) https://artificialintelligenceact.eu/the-act/

(9) https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

(10) https://www.iso.org/standard/81230.html

(11) https://www.nist.gov/itl/ai-risk-management-framework