Cybersecurity and Information Security Newsletter

Issue 24 | February 10, 2023

Table of Contents

Daniel Shin, research scientist with the Commonwealth Cyber Initiative (CCI) Coastal Virginia region, wants to hear from you! Submit any cybersecurity and information security news items or request related topics, via e-mail to dshin01@wm.edu.

This newsletter supports the mission of CCI. To learn more about CCI, including upcoming events, funded research, and news, please visit cyberinitiative.org.

NIST publishes the AI Risk Management Framework

On January 26, 2023, the National Institute of Standards and Technology (NIST) published the *AI Risk Management Framework* (AI RMF), which aims to provide a resource to manage the risks raised by AI systems and promote trustworthiness and responsible development of AI systems. *AI Risk Management Framework [NIST]*, available here. The AI RMF was developed in collaboration with public and private stakeholders through a Request for Information, public comments on AI RMF drafts, and public workshops.

Background

Although AI technologies and traditional software share some risk profiles, AI systems introduce novel risks that are not adequately addressed in traditional software risk frameworks. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* at 43, available here. Risks from traditional software stem largely from programming code and algorithms used by the software. In contrast, most AI systems are largely dependent on training data—instead of the underlying pre-programmed code—for their performance because they utilize Machine Learning algorithms. Machine Learning algorithms generally examine and learn from a given data set (e.g., tables of data points, a corpus of text, and photos) and perform tasks based on their data training sessions. *See Artificial Intelligence: Adversarial Machine Learning [NIST]*, available here. In a Machine Learning-based system, the training data influences how the system acts during deployment, and in some instances, new data collected during system operation is incorporated into the operating data originally created by the training data. As such, a comprehensive review of AI risks must include risks arising out of training and other data.

AI RMF Overview

The National Artificial Intelligence Initiative Act of 2020 (NAII) sets out the U.S. government's coordinated approach to maintain U.S. leadership in artificial intelligence research and lead in the development and use of trustworthy artificial intelligence systems in both public and private sectors. *William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Division E—National Artificial Intelligence Initiative Act of 2020* at 1139, available here. The legislation mandated the Director of NIST to develop a voluntary risk management framework for trustworthy artificial intelligence systems. *15 U.S.C. § 278h–1(c)*, available here. In response to NAII's requirements, NIST began the process of developing the AI RMF.

The goal of the AI RMF is to help minimize potential negative impacts of AI systems—especially in relation to civil liberties and human rights—while facilitating positive impacts (e.g., facilitating innovation), which may lead to more trustworthy and responsible AI systems. *AI RMF 1.0*, *supra* at 9. Trustworthy and Responsible AI refers to AI systems that embed in their design and performance the following characteristics: accuracy, explainability and interpretability, privacy, reliability, robustness, safety, security (resilience), and mitigation of harmful bias. *See Trustworthy and Responsible AI [NIST]*, available here. Incorporating diverse perspectives, disciplines, professions, and experiences as part of risk management can enhance AI trustworthiness, which may, in turn, reduce negative externalities. *AI RMF 1.0*, *supra* at 10-12.

To reach outcomes and actions that enable successful AI risk management and the development of trustworthy AI systems, the Core of the AI RMF presents four functions that

organizations may implement to address risks from AI systems from a lifecycle management perspective.

Figure 1: Graphic representation of the four functions of the AI RMF Core

The **Map** function establishes the context to frame risk related to an AI so that context is recognized and risks related to context are identified. Because risks are not always visible across different AI lifecycles, organizations must ensure that key stakeholders have the proper understanding of AI systems within different contexts.

The **Measure** function uses tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risks and other related impacts.

The **Manage** function refers to "allocating risk resources to mapped and measured risks on a regular basis and as defined by the Govern function." *Id.*, *supra* at 31. After risks are prioritized, they are addressed based on a projected impact on the AI system.

The **Govern** function refers to governance that is infused throughout the AI risk management cycle that places structures, systems, processes, and teams to develop a purpose-driven culture focused on risk understanding and management.

As a companion resource to the AI RMF, NIST also published the NIST AI RMF Playbook, AI RMF Explainer Video, an AI RMF Roadmap, AI RMF Crosswalk, and Perspectives to supplement the AI RMF.

Analysis

The AI RMF is the U.S. government's latest effort to provide guidance on AI technology uses and strategies to minimize risks. For example, in October 2022, the White House's Office of Science and Technology Policy published the *Blueprint for an AI Bill of Rights* (AI Bill of Rights) that devised five principles that "guide the design, use, and deployment of automated systems to protect the . . . public in the age of artificial intelligence." *Blueprint for an AI Bill of Rights [White House's OSTP]*, available here. The AI RMF and the AI Bill of Rights encourage

an interdisciplinary approach to mitigate harm from systems relying on automated or AI technologies. Furthermore, they focus on protecting civil rights and democratic values without overly discouraging innovation or implementation of AI or other technological development.

NIST's AI RMF is the product of collaboration among government representatives, academics, and industry, resulting in a thoughtful, forward-thinking approach to identifying and mitigating AI risks. This voluntary standard may encourage AI development to consider carefully relevant risks, which may help create norms of pursuing the trustworthy standard before deploying AI technologies.

Although the AI RMF is a comprehensive tool, the framework places primary focus on predictive AI, where AI systems are tasked primarily to make a judgment or determination based on a given circumstance. It does not address risks pertaining to generative AI, such as ChatGPT and DALL-E, where the AI system is developed primarily to synthesize data based on its data training set. *AI RMF 1.0*, *supra* at 39 ("[G]uidance available before publication of this AI RMF does not comprehensively address many AI risks. For example, existing frameworks and guidance are unable to . . . confront the challenging risks related to generative AI.").

Given the threats posed by deepfakes and the recent concerns arising out of ChatGPT, NIST has a timely opportunity to study the unique risk profiles of generative AI technologies and supplement its findings on the next version of the AI RMF.

Lawsuits commenced against Stability AI for intellectual property infringement

On January 13, 2023, three full-time artists (Sarah Andersen, Kelly McKernan, and Karla Ortiz) filed a class action lawsuit against Stability AI, Midjourney, and DeviantArt for alleged copyright infringement and other violations arising from the companies' text-to-image synthesis tools and services. *Artists Slam AI Apps As '21st-Century Collage Tool' In IP Suit*, available here; *Implications of AI art lawsuits for copyright laws*, available here. Also, on February 6, 2023, Getty Images, a visual imagery company focusing primarily on stock photos, filed a similar lawsuit against Stability AI in the U.S. District Court of Delaware. *Complaint: Getty Images v. Stability AI*, available here. The company announced previously its intention to commence legal proceedings against Stability AI for intellectual property infringement in the High Court of Justice in London, United Kingdom. *Getty Images Statement*, available here.

Stable Diffusion Overview

Stability AI is a UK-based company developing artificial intelligence-based tools, most notably Stable Diffusion. Stable Diffusion is a text-to-image synthesis Diffusion-based model that was developed using a large data set of text-to-image pairs (i.e., text describing the content of the paired image) from a wide variety of online sources, including copyrighted media. *See Stable Diffusion Public Release*, available here. It synthesizes images based on the user-provided prompts, which are processed to produce complex images, including photorealistic images, with ease. Because Stable Diffusion was trained on copyrighted images, it is possible to

prompt the AI model to synthesize media that take characteristics of a particular visual artist, including those whose works may have been used to train the model. In turn, users are able to produce media that imitate the style of certain artists with similar levels of detail in a matter of seconds. Stable Diffusion was released by Stability AI as a free and open-source AI model, allowing AI enthusiasts to download the model and synthesize images on their own computers. *Stability AI General FAQ*, available here.

Regarding the copyright status of synthesized images, Stable Diffusion only notes that "[t]he area of AI-generated images and copyright is complex and will vary from jurisdiction to jurisdiction." *Id.*

Stable Diffusion is used by DreamStudio (a web-based text-to-image synthesis service launched by Stability AI) and DreamUp (a web-based text-to-image service launched by DeviantArt, a social network focusing on users showcasing visual art media).

Analysis

Both the full-time visual artists and Getty Images allege violations of copyright because Stable Diffusion was trained on copyrighted images for image synthesis purposes without proper authorization by the right holders. Specifically, according to the artists' complaint, Stability AI paid a German-based nonprofit organization LAION (Large-Scale Artificial Intelligence Open Network) to produce a large data set of individual hyperlink references to 5.85 billion images—which included copyrighted works—to train and develop Stable Diffusion. *Class Action Complaint: Andersen v. Stability AI*, available here. As such, the artists argue that any media synthesized by Stable Diffusion would inevitably be derived from copyrighted images. Also, LAION's dataset only contained hyperlinks to images, requiring Stability AI to download the images before training the Stable Diffusion model. *See FAQ [LAION]*, available here ("Any researcher using the datasets must reconstruct the images data by downloading the subset they are interested in.").

In response to the artists' lawsuit complaint, a Stability AI representative suggested that all alleged conduct surrounding Stable Diffusion's development falls under "fair use." Fair use is a legal doctrine that allows the unlicensed use of copyrighted works under certain circumstances. *U.S. Copyright Office Fair Use Index*, available here. To assess a fair use claim, courts evaluate whether an alleged act is a "transformative work," which refers to the alteration from the original work with "new expression, meaning, or message." *Perfect 10 v. Amazon.com, 508 F.3d 1146* at 1164-65, available here. Specifically, a work is transformative if an individual modifies a copyrighted work or uses a copyrighted work in a different context such that the work is transformed into a new creation. *Id.* For example, in *Perfect 10 v. Amazon.com*, the Ninth Circuit Court of Appeals found that Google's use of copyrighted images in its Google Image search engine fell under fair use because copyrighted images were transformed into a pointer to direct users to image search results. *Id.* at 1165.

In a potential legal response, Stability AI would likely argue that the processing of copyrighted images for Stable Diffusion should be transformative work and under fair use. The Stable Diffusion model file does not contain any image data, and the primary purpose of Stable Diffusion is the text prompt-based image synthesis, not a reproduction of training data. Neither the full-time artists' complaint nor the Getty Images' complaint address the potential fair use issue (only the full-time artists' complaint has noted fair use as an anticipatory defense against

the complaint's claims.) *Class Action Complaint: Andersen v. Stability AI*, *supra* at 11 ("Anticipated Defenses . . . Whether any affirmative defense excuses Defendants' conduct, including but not limited to whether some or all of Defendants' conduct is allowed under the Fair Use Doctrine.").

The application of the Fair Use Doctrine is a matter of great relevance. The U.S. Supreme Court is currently considering reducing the scope of the Fair Use Doctrine when evaluating the transformative nature of a given work. *Andy Warhol Foundation for the Visual Arts v. Goldsmith [U.S. Supreme Court Docket], 19-2420*, available here. Specifically, it is contemplating whether the fair use analysis should *not* involve comparing a work's meaning or message from its source material. This case does not specifically address the use of copyrighted materials for AI training purposes. However, if the Court rules to constrain the fair use analysis, Stability AI's legal defense may be impacted and likely deter other generative AI developments from using copyrighted works as part of their training data set.

Even if Stability AI mounts a viable legal defense against copyright infringement, the issue will continue with respect to the relationship between generative AI and artists and writers who created the underlying works the AI is trained on. Even if the Fair Use Doctrine allows for the appropriation of copyrighted works for synthesizing new media, such technologies may threaten the ecosystem of original, creative expressions. By developing tools that can automate the synthesis of artistic media, the economic environment supporting visual artists would likely get threatened as visually stunning AI-generated imagery gets flooded online. Eventually, the mass production of synthesized media may crowd out the demand for authentic media due to the low cost of media synthesis. Policymakers may have a timely opportunity to mitigate the harm from the market disruptions arising from generative AI without overly dissuading innovation in AI development.