Protecting Against the Legal Risk of AI-Enabled Copyright Infringement

ADVISORY **Publish Date** July 09, 2024 **By** RANE Cyber Compliance

Since the advent of the Artificial Intelligence (AI) boom in late 2022, companies across the globe have been integrating AI for various purposes, creating, along with immense benefits to cost and efficiency, a number of risks, including that of potential inadvertent copyright infringement. A wide range of companies that use AI to produce marketing materials, advertising, or other content could potentially be at risk, particularly if their tools produce outputs that are highly similar to existing, copyright-protected works. To help understand the risks of copyright infringement from AI-generated materials, **RANE** spoke to **Joseph DeMarco,** an expert on law relating to emerging technologies and partner at **DeMarco Law, PLLC**, a litigation and counseling boutique dedicated to the protection of intellectual property, emerging e-commerce and internet law, and information privacy and security.

DeMarco tells **RANE** that AI models are trained on millions of pieces of data, typically including images, videos or audio clips, all of which may be protected by copyright law. He says "generative AI uses all of this underlying data to create new works or new data in response to prompts from humans." Because AI models are built by scraping the internet of large swaths of information in the public domain, many copyright-protected works are likely included in the training data of any large language model. As such, when generative AI models create various types of content, including written works, images, videos or music, the results are based on a conglomeration of authentic material from the internet. Because of this, AI-generated content often mimics some aspects of existing works or, in some cases, entire works as a whole.

Current Legal Cases

The use of copyrighted material in these models is prolific because, as **DeMarco** says, "copyright law may protect something as simple as a photo that you take on your iPhone, even if you do not register it with the copyright office." While this risk is in its nascency, a number of lawsuits have already been filed, largely against AI companies over fair use of copyrighted material in their products, including more than a dozen lawsuits against AI company and ChatGPT creator OpenAI.

- A flurry of news organizations, including The New York Times, The Intercept, Raw Story and AlterNet Media Inc., are among those that have brought lawsuits against OpenAI for copyright infringement in using their news stories and content to train its AI models. Writers Sarah Silverman, Christopher Golden and Richard Kadrey have also filed complaints against OpenAI for unlawfully scraping their works from the internet. In February 2024, a federal judge rejected most of the authors' lawsuits, saying that the plaintiffs failed "to explain what the outputs entail or allege that any particular output is substantially similar – or similar at all – to their books." Nonetheless, the judge declined to dismiss a claim of unfair competition.

- In February 2023, Getty Images filed a lawsuit against Stability AI, accusing the company of infringing more than 12 million photographs, their associated captions and metadata to build its Stable Diffusion and DreamStudio offerings. Along with copyright infringement, the case also accuses Stability AI of trademark infringement due to its technology's ability to replicate Getty Images watermarks in the models' outputs.

- In the case of Anderson vs. Stability AI, visual artists in January 2023 filed a putative class action against Stability AI alleging direct and induced copyright infringement along with a host of other claims, including Digital Millennium Copyright Act violations, false endorsement and trade dress claims based on the creation and functionality of Stability AI's Stable Diffusion and DreamStudio as well as on two separate companies – Midjourney Inc.'s generative AI tool and DeviantArt's DreamUp. In this case, artists Sarah Anderson, Kelly McKernan and Karla Ortiz claimed that Stability AI downloaded or acquired billions of copyrighted images without permission to be used as training images for a variety of generative AI platforms to produce output images in their particular artistic styles. A district court judge has since thrown out all but one of the claims that the artists asserted, asking for more specifics on how each defendant was involved in the claimed infringement, which could indicate a higher tolerance in the courts for AI-generated content based on copyrighted material going forward. An interesting aspect of this case is the allegations against Midjourney, as the court has said plaintiffs need to clarify whether the allegations were based on Midjourney's use of Stable Diffusion or on Midjourney's own independent use of training images for its own product. If the former, it could indicate how lawsuits can be brought against firms that use external AI systems trained on copyrighted materials.

How Does AI Compare to Copyright in Previous Technology Waves

Similar copyright issues have previously arisen with the emergence of novel technologies. **DeMarco** gives the example of radio, saying that at the dawn of radio, stations were playing copyright-protected music generated by artists. **DeMarco** says that this issue was solved through a royalty system that allows artists to be compensated based on formulas and aggregated data when radio stations play their music. A more modern example would be streaming platforms such as Spotify or sites such as Netflix or Hulu, which have since been granted fair use privileges to stream copyrighted songs, movies and shows. This could potentially be mimicked going forward in AI cases, as there have already been instances of news organizations settling with AI companies for the use of their works. For example, in July 2023, AP News made a deal with OpenAI to license AP's archive of news stories, and prior to the New York Times lawsuit against OpenAI, the company spent months in negotiations attempting to reach a similar settlement. As such, much like the examples of radio and modern streaming services, **DeMarco** tells **RANE** that the issue boils down to fair use and "whether or not the AI program is fairly using that underlying copyrighted work." However, a key distinction between these issues and AI copyright issues is that in the case of radio or streaming services, it is an existing copyrighted work that itself is being replayed. **DeMarco** says that "there the issue is a lot cleaner from a legal point of view," compared to the case of AI, "where you have a million songs being fed into a data tool which then spits out a song [or piece of content] which is not an exact copy of any of the million songs [or other copyrighted works] that were put into the tool, but still created drawn upon those songs." Another potential difference from previous copyright issues is that AI has the ability to produce works that are used for vastly different purposes than the original work and thus may not impact the original market.

Where is This Issue Going?

DeMarco tells **RANE** that these initial cases brought against AI companies for copyright infringement are still in their early stages and that "we are not going to have hard and fast law until it reaches the Appellate level and maybe not until it reaches the U.S. Supreme Court level," which he says could take years. **DeMarco** says that courts are in the process of deciding now "whether or not fair use doctrines will apply to protect the AI-generated content," along with the question of who owns the outputs of the AI programs, which may depend on and use a broad array of copyright-protected works. **DeMarco** says this can be a complicated issue: "fair use is a kind of squishy doctrine, and it looks to a number of factors to determine whether or not that kind of situation should be allowed or prohibited under the copyright law." The fair use test, he says, is "a kitchen sink test," which refers to a programming term derived from the phrase "everything but the kitchen sink," which means almost anything one can think of. When asked about this aspect of manipulation, which sets AI copyright apart from other previous emerging technology cases, **DeMarco** says that "the courts will get at those issues because they will be looking at the purpose and character of the use of the work, the nature of the copyrighted work, how similar the output of the AI tool is to the original work, the effect on the potential market for the original work, as well as the purpose for which it was intended." He says that while "the existing legal regime of copyright law can get you into court, the question is how the courts are going to apply this very loose doctrine of fair use to either allow those cases to go forward or to ding those cases, and that remains to be seen."

Risks for Organizations:

Though these cases have yet to be settled and their outcomes remain unclear, a larger issue will likely emerge beyond fair use of copyrighted materials in training datasets to encompass fair use of AI-generated content for business purposes, which may implicitly include copyrighted material. This would impact companies that use AI to generate marketing materials, press releases or other content creation, for example.**DeMarco** says that "any company that uses AI tools to create content is potentially at risk." Particularly at risk are marketing, advertising or communications firms, which may rely more heavily on AI tools for business functions, but **DeMarco** says, "really any company that may even in-house produce marketing tools, sales tools or promotional tools could be at risk." He goes on to give an example saying, "if a company engages an AI tool to create music and images for a marketing campaign that winds up infringing on copyrighted works, the original author could potentially sue both the AI developer and the client company that used that tool." In several cases that have been brought against AI companies, including the Anderson vs. Stability AI and the Sarah Silverman-led suit against OpenAI, judges have said that there is not sufficient evidence to show that the outputs violate copyright laws.

While the issue of fair use and competition have yet to be settled in these cases, something that could more immediately pose risks for organizations is if they fail to vet the output of generative AI services to ensure that models are not producing works that are markedly similar to existing copyright protected works. While these cases have not gained traction due to claims that every output violated copyright law, it is more likely that as companies use the AI models more, there will be a few outputs that could mirror copyrighted works. While the vast majority of outputs may not verbatim mimic written works or look or sound extremely similar to artist content, studies have found that in some cases, generative AI will produce works that are extremely similar to existing, protected materials, especially when users give these models certain specific prompts. These more infrequent cases of extremely similar outputs are the instances that are more likely to cause problems for organizations rather than the bits and pieces and strings of existing materials baked into the majority of generative AI outputs.

Additionally, there is more likely to be a case brought if the AI-generated materials impact the market of the original works that went into them, such as in the case of AI-generated music, which mimics artist vocals or styles and can thus compete with their authentic works. This is why, in the case of the Sarah Silverman-led lawsuit against OpenAI, the main outstanding question of the case is the competition aspect. **DeMarco** tells **RANE** that there is a much looser case for instances in which an "AI-generated work is used for a completely different purpose for anything that went into it and exists to satisfy a completely different market, and therefore does not affect the sales of the original works."

Best Practices

To best protect themselves from these legal risks, **DeMarco** recommends that, first and foremost, companies and their employees be aware of these issues. Awareness of the copyright risks that organizations face will help them to better proactively prepare for such contingencies. Beyond understanding the baseline risks, **DeMarco** also recommends that organizations understand the tools already in use throughout their organization. He says it is imperative for organizations to vet every AI tool they use, saying that "the most important thing is to do an AI audit to make sure that you fully understand the AI tools your company is using." His advice to firms is to "make sure that you understand what protections the AI tools that you are using have put in place to prevent their use of infringing copyright-protected works." Beyond the tools available within an organization, it is also important to know and understand which AI tools a company's vendors may use. Additionally, according to **DeMarco,** organizations should also make sure to "exclude copyrighted materials from anything you have developed or any of your providers have developed." He also recommends that organizations "think about your contracts with those organizations," and for contracts with AI companies, "make sure that the agreements memorialize that the data used to train the models has been properly licensed or is in the public domain." This way, **DeMarco** says, "you should get indemnification from those companies if anyone sues you in addition to them for copyright infringement."

About the Expert:

Joseph DeMarco is a partner at DeMarco Law and a leading expert in internet crime and law relating to emerging technologies. DeMarco previously served as Assistant United States Attorney for the Southern District of New York, where he founded and headed the Computer Hacking and Intellectual Property Program, a group of five prosecutors dedicated to investigating and prosecuting violations of federal cybercrime laws and intellectual property offenses. Prior to joining the United States Attorney's Office, DeMarco was a litigation associate at Cravath, Swaine & Moore in New York City, where he concentrated on intellectual property, antitrust, and securities law issues for various high-technology clients. Since 2002, DeMarco has also served as an Adjunct Professor at Columbia Law School, where he teaches the upper-class Internet and Computer Crimes seminar. He has spoken throughout the world about cybercrime, e-commerce and IP enforcement. He has lectured on the subject of cybercrime at Harvard Law School, the Practicing Law Institute, the National Advocacy Center, and the FBI Academy in Quantico, Virginia. He has served as an instructor on cybercrime to judges attending the New York State Judicial Institute.