1. **As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?**

Generative AI technology, with its ability to autonomously create copyrightable content, presents both promise and peril. On one hand, it unlocks new avenues for creativity, streamlining content production and allowing artists to explore uncharted realms of imagination. This has the potential to catalyze innovation and foster diversity within creative industries. Collaborations that blend human ingenuity with AI-generated elements can lead to unexpected and groundbreaking outcomes, pushing the boundaries of artistic expression.

However, generative AI also introduces a range of challenges and risks. Age-old concepts like authorship and copyright are thrown into disarray as it becomes increasingly difficult to determine the true creator of AI-generated content. An illustrative example is the recent case of an AI-generated song titled "Heart on My Sleeve," featuring renditions of Drake and The Weeknd's voices, which was submitted for Grammy consideration despite neither artist's involvement in its creation. The song is the brainchild of a mysterious entity known as Ghostwriter. Recording Academy CEO Harvey Mason Jr. has argued that the submission qualifies for Grammy consideration because the lyrics were written by a human, even though the vocal performances are computer-generated (Variety Magazine). The emergence of AI-generated music has given rise to concerns and legal issues within the music industry. Universal Music Group, representing Drake and The Weeknd, invoked copyright violations to remove the song

from platforms (NPR). This controversy has ignited discussions about artists' rights, copyright, and the necessity of safeguarding human creative expression in the era of AI-generated content.

Furthermore, concerns arise regarding the quality and originality of AI-generated works. While AI can produce content rapidly, it may flood the market with derivative or subpar material, potentially diminishing the value of human-created content.

The impact of generative AI extends to various stakeholders. Creators can benefit from AI as a potent creative tool but must grapple with issues of authorship and copyright when AI plays a significant role in content production. Copyright owners face the challenge of protecting their rights in AI-generated content, particularly when AI is a central element of the creative process. Technology developers tread carefully, navigating complex copyright laws and ethical considerations while delivering innovative AI solutions. Researchers harness AI's potential for data analysis and content generation but must exercise vigilance regarding its responsible use. The public gains increased access to AI-generated creative works but requires transparent labeling to distinguish between human and AI-generated content, ensuring informed consumption.

Generative AI technology reshapes the creative landscape, offering new opportunities while raising profound questions about authorship, copyright, and creative authenticity. As this technology continues to evolve, it calls for ongoing discussions, robust policy development, and ethical frameworks to harness its creative potential while addressing the challenges it poses to creators, copyright owners, developers, researchers, and the broader public.

6. **What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?**

In the field of training AI models, a diverse array of copyright-protected materials is harnessed for the purpose of imparting knowledge and facilitating pattern recognition. These materials encompass a wide spectrum, including but not limited to photos, paintings, books, computer source code, and other forms of creative content. As outlined in an article by intellectual property law attorney Diana Bikbaeva, these datasets may inadvertently incorporate copyrighted materials without the knowledge or consent of the copyright owners, which has sparked concerns in the legal and AI communities.

According to an article Ellen Glover, a senior staff reporter covering artificial intelligence and data science, the process of training AI models typically involves the assembly of vast datasets that serve as the foundation for machine learning. These datasets are instrumental in teaching AI systems to recognize patterns and make predictions independently, reducing reliance on human programming. As highlighted by Bikbaeva, the size and scale of these datasets can be staggering, with examples like LAION-5B, comprising a colossal 5.85 billion image-text pairs. Such datasets, often amassed from publicly available sources, can include copyrighted materials obtained from photos, paintings, books, and other creative works.

The curation of these datasets is a complex task, as pointed out by Glover. In some cases, the inclusion of copyrighted materials may be unintentional, as copyright owners might be unaware of or have not consented to their works being used for machine learning purposes. This issue is particularly pertinent when AI developers tap into open-source code (The Fashion Law). Open-source code, although publicly accessible, may come with licenses that require attribution to the original authors and sharing of derivative works with the public for free. Failure to adhere to these licenses can lead to legal disputes, as exemplified by the GitHub Copilot lawsuit, where GitHub and Microsoft were alleged to have violated open-source code licenses.

Moreover, the inclusion of copyrighted materials in AI training datasets raises fundamental questions about the intersection of copyright law and AI development, as emphasized by both articles. According to Bikbaeva, copyright law grants copyright holders six exclusive rights, including the right to make copies of the work and prepare derivative works. Machine learning often involves making copies of copyrighted materials, such as when assembling training datasets. Furthermore, if the output data generated by AI closely resembles copyrighted materials from the training dataset, it implicates the right to create derivative works, which can potentially result in copyright infringement claims unless exceptions like the "fair use" doctrine apply.

In the context of fair use, Bikbaeva explores the criteria used to determine whether the use of copyrighted materials in AI training datasets qualifies as fair use. Factors such as the purpose and character of use, the nature of the copyrighted work, the amount and substantiality of the portion used, and the effect on the potential market for the copyrighted work are considered. Specifically, a transformative use that serves a different purpose from the original work is more likely to be deemed fair use, as it expands public knowledge and understanding without negatively impacting the market for the original work.

As for the commonly used AI software ChatGPT-3 developed by OpenAI, it was trained on a diverse corpus of text from the internet. The specifics of the dataset, including the sources, were not disclosed publicly by OpenAI. However, the training process generally involves using large-scale datasets to teach the model to predict the next word in a sentence, a process known as unsupervised learning. The model learns language patterns, grammar, and context from this data. While the training data includes text from various sources, it is crucial to note that ChatGPT-3's

responses are generated based on patterns learned from this data and do not involve the direct copying of copyrighted materials.

The use of copyright-protected training materials in AI development is a complex issue, encompassing various types of creative content and legal considerations. The curation of datasets, the inadvertent inclusion of copyrighted materials, and the application of fair use principles all play significant roles in shaping the landscape of AI development in relation to copyright law. Addressing these issues will require careful consideration of the rights of copyright holders, the transformative nature of AI applications, and the evolving legal landscape surrounding AI and intellectual property.

8. **Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question.**

Determining whether the unauthorized use of copyrighted works to train AI models constitutes fair use is a complex matter that hinges on various factors, including legal precedents and the specifics of each case. Fair use is a doctrine within copyright law that permits limited use of copyrighted materials without the copyright owner's consent for purposes such as criticism, comment, news reporting, teaching, scholarship, or research. However, applying this doctrine to AI training data introduces nuances and uncertainties. One critical factor in assessing fair use is the "purpose and character of use." Courts examine whether the use of copyrighted materials is transformative, meaning it adds new elements or serves a different purpose compared to the original work. This transformative nature can weigh in favor of fair use. An example of this principle can be found in the case of Sega Enterprises Ltd. v. Accolade, Inc. (1992), where

copying a competitor's computer program code to understand its unprotected functional elements and ensure compatibility with a gaming console was considered fair use because it had a transformative purpose.

The "nature of the copyrighted work" is another factor. Courts differentiate between more factual and more creative works. Fair use is typically more likely when the copyrighted work leans towards the factual side, as copyright protection primarily applies to creative expression. For AI training, this means that using copyrighted materials that are highly creative may present more challenges in establishing fair use.

The "amount and substantiality of use" is also assessed. Courts consider how much of the copyrighted material is used in relation to the work as a whole. Using only a small portion for a specific purpose may lean toward fair use, while extensive use could weigh against it. The principle behind this factor was evident in Authors Guild v. Google, Inc. (2015), where Google's creation of thumbnails or snippets of copyrighted books was deemed transformative fair use because it served a different function than the original creative content.

Finally, the "effect on the potential market" is a crucial consideration. Courts determine whether the unauthorized use of copyrighted materials negatively impacts the market for the original work. If it serves as a substitute for the original or competes directly with it, it may hinder fair use. The goal is to strike a balance between the rights of the copyright owner and the benefit to the public. However, there is limited case law directly addressing the use of copyrighted works in AI training, which leaves room for uncertainty. One notable case that provides some guidance is the ongoing GitHub Copilot lawsuit. While not primarily a fair use case, it underscores the complexities of using open-source code in AI training. GitHub and Microsoft were sued for allegedly violating open-source code licenses, emphasizing the

importance of respecting licensing agreements when incorporating copyrighted materials into AI models. This case highlights the need for clarity in navigating the intersection of AI development and copyright law.

Determining when the unauthorized use of copyrighted works for AI training constitutes fair use is a nuanced and evolving legal issue. Factors such as the transformative nature of the use, the nature of the copyrighted work, the amount used, and the impact on the market all come into play. As AI development continues to progress, it is likely that more legal precedents will emerge to provide greater clarity in this complex area of law. Until then, each case must be considered on its own merits, making it imperative for AI developers to tread carefully and seek legal guidance when using copyrighted materials for training purposes.

Works Cited

"Artificial Intelligence and Copyright." Wikipedia, Wikimedia Foundation, 13 Oct. 2023, en.wikipedia.org/wiki/Artificial_intelligence_and_copyright.

Bikbaeva, Diana. "Ai Trained on Copyrighted Works: When Is It Fair Use?" The Fashion Law, 10 May 2023, www.thefashionlaw.com/ai-trained-on-copyrighted-works-when-is-it-fair-use/.

Glover, Ellen. "Ai-Generated Content and Copyright Law: What We Know." Built In, 23 Aug. 2023, builtin.com/artificial-intelligence/ai-copyright.

Lang, Courtney. "Current AI Copyright Cases – Part 1." Copyright Alliance, 24 May 2023, copyrightalliance.org/current-ai-copyright-cases-part-1/.

Shanfeld, Ethan. "Ghostwriter's 'heart on My Sleeve,' the AI-Generated Song Mimicking Drake and the Weeknd, Submitted for Grammys." Variety, Variety, 6 Sept. 2023,

variety.com/2023/music/news/ai-generated-drake-the-weeknd-song-submitted-for-grammys-1235714805/.

“Timeline of Artificial Intelligence.” Wikipedia, Wikimedia Foundation, 15 Oct. 2023, en.wikipedia.org/wiki/Timeline_of_artificial_intelligence.

Veltman, Chloe. “When You Realize Your Favorite New Song Was Written and Performed by ... Ai.” NPR, NPR, 21 Apr. 2023, www.npr.org/2023/04/21/1171032649/ai-music-heart-on-my-sleeve-drake-the-weeknd.