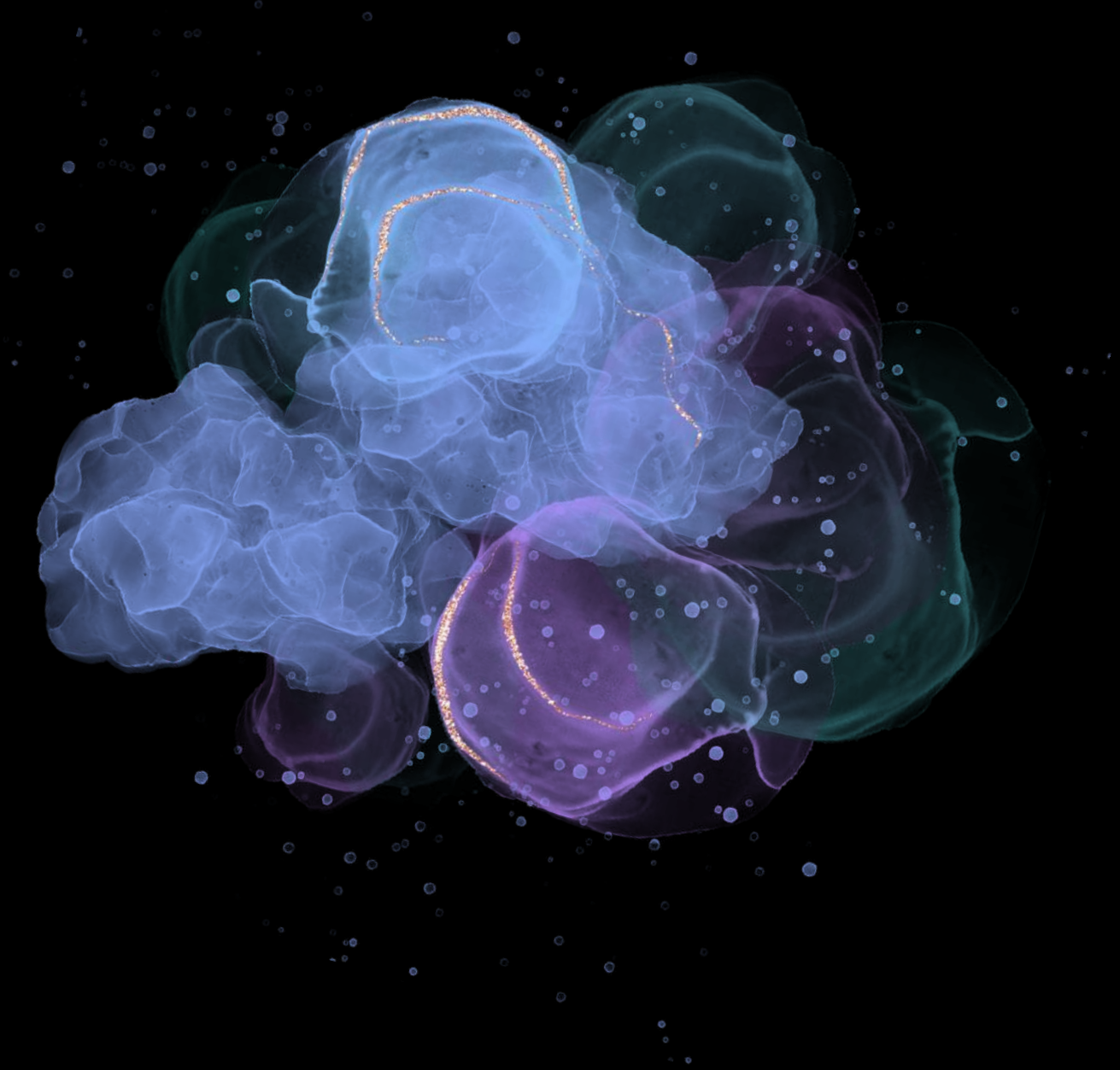
SeaQVN Suggests

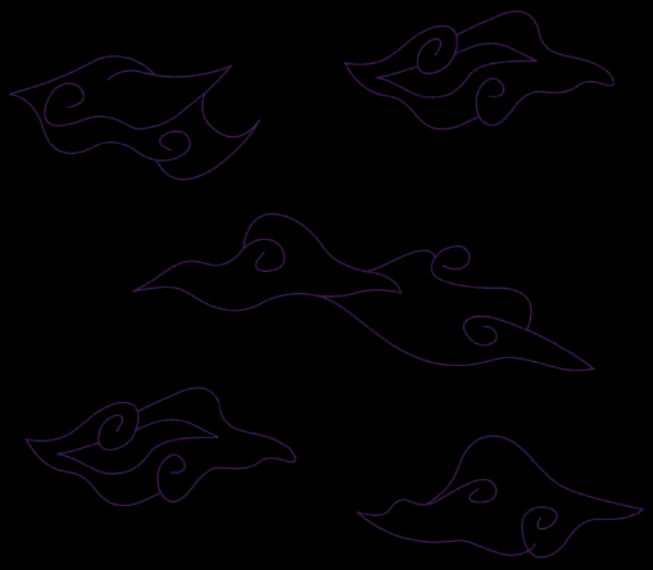
A response to Artificial Intelligence and Copyright Notice by the U.S. Copyright Office



Sept' 2023

Link to the Notice:
https://www.regulations.gov/document/COLC-2023-0006-0001

Questions raised in the Notice and SeaQVN's responses



Responses have been generated as a human and AI collaboration: The generated information was meticulously screened for its integrity and factual accuracy to the best capabilities of the human author.

FOREWARD

To the esteemed members of the U.S. Copyright Office, and to the broader community engaged in the dynamic intersection of artificial intelligence and copyright law,

The rapidly evolving landscape of artificial intelligence (AI) has thrust us into a new paradigm. As AI systems demonstrate capabilities ranging from creating art and composing music to authoring prose and generating code, traditional notions of creativity and authorship are being profoundly reshaped. Against this backdrop, we present "SeaQVN Suggests," a meticulous response to the U.S. Copyright Office's probing inquiry into the relationship between AI and copyright.

Central to our response is an unwavering commitment to a broad spectrum of stakeholders. From the innovators and developers at the helm of AI breakthroughs to the artists, writers, and everyday creators harnessing these technologies. From the consumers delighting in AI-generated content to the third parties and regulators diligently working to maintain a delicate equilibrium between innovation and protection.

SeaQVN's core motivation in this endeavor is clear and unwavering: to nurture an ecosystem that upholds the best interests of everyone involved. In our journey through the AI domain, we've become both participant and advocate. Recognizing the transformative power AI wields, we also understand the urgent need for a robust, equitable, and forward-looking framework to guide its integration into our creative landscapes.

Our report aims to provide a rich tapestry of insights, recommendations, and reflections. While we delve into the granular details of AI's operational mechanisms, generative content, and the conundrums of attribution and accountability, our overarching vision remains steadfast: to champion a path that acknowledges the sanctity of original creation, yet embraces the revolutionary impact of AI.

As the U.S. Copyright Office sifts through these insights and perspectives, it is our hope that they will serve as both a resource and a catalyst. Whether dissecting the significance of individual contributions to AI-generated outputs or exploring the intricacies of sections like 17 U.S.C. 1202(b), our aim is to offer a comprehensive view that assists in your vital decision-making processes.

To the developers, creators, consumers, and all who interact with the marvel that is AI, may "SeaQVN Suggests" resonate as both reflection and guidance. A reflection of our current, complex intersection of AI and copyright, and guidance towards a harmonious future where both creativity and innovation flourish side by side.

In presenting this report, we, at SeaQVN, extend our deepest gratitude to the countless voices and visions that have influenced our stance. Your experiences, concerns, aspirations, and insights have been instrumental. It's with a spirit of collaboration, mutual respect, and hope for the future that we put forth "SeaQVN Suggests."
May our combined endeavors illuminate the path to a future where AI and creativity coalesce, prosper, and inspire.

Warm Regards,
Nitish Arora
Founder & CEO, SeaQVN

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?

The potential of generative AI systems to produce material that would be copyrightable if created by a human author is a groundbreaking and controversial development in the intersection of technology and intellectual property. This capability opens up a myriad of possibilities, as well as challenges. Here are my views on the potential benefits and risks, along with the current and prospective impact on various stakeholders:

Benefits:

Scalability and Efficiency: Generative AI can produce vast amounts of content in short periods, aiding in sectors like entertainment, research, and design where rapid content generation is invaluable.

Bridging Creative Gaps: For creators who experience creative blockages, AI can suggest ideas, patterns, or compositions, serving as a collaborative tool.

Customization: AI can tailor content for individual users, providing personalized experiences in fields like marketing, entertainment, and education.

Accessibility: Generative AI can help in producing content in multiple languages or formats, making information and entertainment more accessible.

Risks:

Originality and Authenticity: As AI-generated content becomes more common, discerning original human-made content can become challenging. The "human touch" or emotional depth in creative works might be diluted.

Economic Impact on Creators: If AI-generated content becomes a cheaper and faster alternative, human artists, musicians, writers, etc., might face economic challenges and job insecurity.

Legal Challenges: The blurred lines between AI-generated content and human-made content can lead to copyright disputes. Determining ownership, rights, and royalties can become complex.

Ethical Concerns: AI might generate content that is inappropriate, offensive, or misleading. There's also a risk of the technology being used for malicious intent, such as deepfakes.

Impact on Stakeholders:

Creators: While some may benefit from using AI as a supplementary tool, others might face competition, especially if their work can be easily and closely replicated by AI.

Copyright Owners: The emergence of AI-generated content challenges traditional notions of copyright. Ownership and rights for AI-generated content remain gray areas. This could lead to potential revenue losses or legal battles.

Technology Developers: Those who pioneer and refine generative AI technologies stand to gain significantly, both in terms of economic benefits and influence over the future direction of content creation.

Researchers: The rise of generative AI presents an exciting domain for research, from refining algorithms to studying the socio-economic impacts of such technology.

The Public: Consumers stand to benefit from a wider variety of content and potentially lower prices. However, they might also struggle with discerning the authenticity of content or face ethical dilemmas related to consuming AI-generated content.

In conclusion, the emergence of generative AI as a tool for content creation is a double-edged sword. While it offers remarkable possibilities for enhancing creativity and content accessibility, it also raises profound questions about the nature of originality, the value of human creativity, and the potential socio-economic ramifications of widespread AI adoption in the creative sphere. As with most disruptive technologies, a balanced approach, informed by continuous dialogue among stakeholders, will be essential to harness its benefits while mitigating its challenges.

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

Music Industry:

Authenticity: As AI-generated music becomes more prevalent, distinguishing between human and machine creativity can become blurred. It raises questions about the soul, emotion, and authenticity of music.

Royalties and Ownership: If a song becomes a hit and it's AI-generated, who gets the royalties? The developers? The operators of the AI?

Sampling and Influence: AI can generate music influenced by various artists. This could lead to potential copyright infringements if AI unintentionally replicates specific melodies or rhythms.

Film and Television:

Scriptwriting: AI can be used to analyze successful scripts and generate new ones. The originality and rights to such scripts can become controversial.

Deepfakes: AI can generate realistic video footage of real people saying or doing things they never did. This has vast implications for misinformation, consent, and the authenticity of content.

Character and Plot Development: AI might suggest character arcs or plot twists based on popular trends, leading to concerns about the homogenization of content and loss of human touch.

Publishing (Books, Articles, etc.):

Authorship: If an AI writes a book or an article, who is the author? The machine cannot hold copyrights, so does it go to the developer, the operator, or neither?

Plagiarism: AI can unintentionally generate content that mirrors existing work, leading to potential copyright claims.

Quality Control: There's a risk of diluting the quality of content in the industry with mass-produced AI-generated materials.

Visual Arts (Photography, Paintings, etc.):

Originality: AI can generate artworks based on historical styles. Distinguishing between a genuine Picasso and an AI-generated Picasso-style painting can be challenging.

Digital Art Market: With platforms like Artbreeder allowing users to "breed" images, the line between creator and curator is blurred.

Value and Pricing: The value of art is often tied to its originality and the artist's intent. AI-generated art challenges these notions, potentially impacting art market economics.

Software and Gaming:

Procedural Generation: Many games use AI for procedural generation of game environments. It might raise questions about the uniqueness of game assets and potential infringement.

Game Testing and Balancing: AI can simulate player behavior. It could inadvertently lead to mimicking gameplay elements from other games.

Character AI: Deep learning can make game characters more realistic, potentially leading to issues around character rights and likenesses.

Architecture and Design:

Design Authenticity: AI can suggest designs based on popular trends, which might diminish the value of unique human designs.

Safety and Compliance: AI-generated designs must be evaluated rigorously for safety, leading to potential liabilities.

Cultural Sensitivity: AI may not always account for cultural and local nuances, which are essential in architecture and design.

Tech and Innovation:

Patents: If AI suggests or creates a novel solution, who holds the patent? This is especially relevant for sectors relying on innovation.

Data Training Bias: Innovations driven by AI can be skewed if trained on biased data. This can lead to non-inclusive tech solutions.

Speed of Innovation: AI can accelerate the pace of innovation, which could outstrip the ability of regulatory bodies to keep up.

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

1. "Artificial Intelligence — The Revolution Hasn't Happened Yet" by Michael Jordan.

Overview of the AI landscape, touching on its economic, social, and technical implications.

https://hdsr.mitpress.mit.edu/pub/wot7mkc1

2. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation".

Addresses the potential risks of AI, especially in creative tasks and implications for various industries.

https://arxiv.org/abs/1802.07228

3. "Generating Rembrandt: Artificial Intelligence, Copyright, and Accountability in the 3A Era—The Human-Like Authors Are Already Here—A New Model." by Y. Elkin-Koren, N. Weinstock and N. Perel.

Investigates copyright challenges posed by AI creators, particularly in the context of accountability.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046067

4. "Notes from the AI frontier: Modeling the impact of AI on the world economy" by McKinsey & Company.

Broader economic implications of AI across sectors.

https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? How important a factor is international consistency in this area across borders?

South Africa: In 2020, South Africa's patent office granted a patent recognizing Dabus AI as the inventor. The decision was groundbreaking, as it challenged traditional notions of human authorship and inventiveness. However, it's crucial to understand that while Dabus AI was recognized as the inventor, the rights to the invention were still awarded to a human (the AI system's owner). This raises questions about the distinction between "inventorship" and "ownership."

Australia: In a similar vein, Australia's Federal Court ruled in 2021 that Dabus AI could be recognized as the inventor for patent purposes. However, like South Africa, while the AI could be named as the inventor, it didn't possess the rights to the patent; those rights were vested with the human owner.

European Union: The European Union has had discussions about AI and intellectual property. The European Parliament has called for rules that can address AI's challenges in various sectors, including copyright. The EU's stance has generally leaned towards recognizing only humans as the rightful claimants of copyright or patent rights.

United Kingdom: The UK's Intellectual Property Office (UKIPO) traditionally requires a human inventor for patents. They have previously rejected patents where Dabus AI was listed as the inventor, adhering to the idea that only humans can be recognized under the current legal framework.

China: China, a significant player in the AI industry, has been proactive in revising its guidelines concerning AI and intellectual property. The country has proposed guidelines which, while not granting AI entities copyright or patent rights, do look into AI's role in both the creation and infringement of such rights.

Japan: Japan's Patent Office has also been considering AI in the context of intellectual property. They've sought public opinions on issues like whether AI-generated works should be protectable and how to approach cases where AI infringes on intellectual property.

Importance of International Consistency:

International consistency in copyright regulations, especially concerning AI, is crucial for several reasons:

Cross-border Collaborations: Many AI projects and ventures are collaborative efforts spanning multiple countries. Consistent regulations can facilitate smoother collaborations and ensure that intellectual property rights are upheld across borders.

Trade and Commerce: As AI products and solutions are often marketed and sold internationally, having a consistent legal framework can help businesses navigate copyright challenges in different markets.

Legal Recourse: In cases of copyright infringements that involve parties from different countries, a consistent international legal framework can provide clarity and ease in seeking legal remedies.

Innovation and Research: International consistency can provide a stable environment for researchers and innovators, ensuring that they don't inadvertently infringe on copyrights as they develop and test AI solutions across different countries.

Setting Precedents: As AI continues to evolve, new copyright challenges will emerge. Having a consistent international approach can help in setting clear precedents for addressing these challenges.

In conclusion, while each country has its socio-cultural and economic nuances that shape its copyright laws, there's undeniable value in striving for international consistency, especially in the rapidly evolving domain of AI. It not only simplifies cross-border operations but also provides a clearer path for future innovations in the space.

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

The emergence of generative AI systems has undeniably blurred the traditional lines of copyright, bringing forth novel challenges that weren't present during the framing

of the existing copyright legislation. The question of whether new legislation is warranted is a complex one, but here's a breakdown of the key points to consider:

Arguments in Favor of New Legislation:

Clear Definitions: The current copyright law often operates on the assumption that every piece of work has a human creator behind it. A new legislation can provide a clear definition of what constitutes an AI-generated work, offering clarity to creators, developers, and users alike.

Ownership and Attribution: Generative AI creates content without a human's direct input, raising questions about ownership. Should the creator of the AI system be considered the author, or should the AI itself? New laws can help establish clear ownership guidelines.

Fair Use and Licensing: With AI often being trained on vast datasets, many of which include copyrighted material, there's a need for clearer guidelines on what constitutes fair use in the context of AI training and generation.

Economic Incentives: Recognizing AI-generated works might encourage further investment in AI research and development. On the other hand, it can also ensure that human creators whose work is used to train AI systems are adequately compensated.

International Standards: As AI becomes a global phenomenon, there's a need for consistency in how different nations approach AI-generated content, ensuring that there aren’t vastly different copyright standards from one country to the next.

Arguments Against New Legislation:

Intrinsic Value: One could argue that AI-generated works lack the inherent creativity and originality that human-generated content has, and thus might not warrant the same level of protection.

Implementation Challenges: Drawing the line between AI-assisted creations (where humans use AI as a tool) and wholly AI-generated works could be problematic.

Economic Impact: Providing copyright for AI-generated content might flood the market with copyrighted materials, making it harder for human creators to monetize their works.

Technological Evolution: The rapid pace of AI evolution means that any laws framed today might become obsolete in a few years.

Potential Approaches:

Limited Duration: AI-generated works could be given copyright protection, but for a shorter duration than human-created content.

Compulsory Licensing: A system where AI-generated works automatically fall under a licensing regime, allowing for their use under set conditions.

Rights to Train: Clearly define the rights of AI developers to use copyrighted works for training models, potentially establishing a licensing system for such uses.

Attribution Over Ownership: Instead of full copyright, AI-generated content could have a mandatory attribution requirement, ensuring that users know the content is machine-generated.

In conclusion, while there are valid arguments on both sides, there's a growing consensus that the current copyright framework may not be adequately equipped to handle the nuances introduced by generative AI. Whether it's through entirely new legislation or amendments to existing laws, there's a clear need for a more comprehensive framework that addresses the unique challenges posed by AI in the creative domain.

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

The training of AI models, especially deep learning models, requires vast amounts of data. Depending on the intended function of the AI, different copyright-protected training materials might be used. Here’s a breakdown of the kinds of copyright-protected materials commonly used and how they are collected and curated:

Textual Data:

Sources: Books, articles, journals, blogs, websites, and other written content.

Collection: Web scraping tools, APIs (like Twitter API for collecting tweets), digital libraries, and datasets shared by researchers or institutions.

Curation: Cleaning and preprocessing data to remove duplicates, irrelevant information, and format inconsistencies.

Images and Visual Data:

Sources: Photographs, illustrations, paintings, digital art, comics, and more.

Collection: Websites, digital image databases, cameras, or purchasing/licensing from image providers.

Curation: Image augmentation (rotations, flips, cropping, etc.) for increased dataset diversity, and categorization based on content or features.

Audio Data:

Sources: Music tracks, podcasts, radio shows, voice recordings.

Collection: Audio streaming platforms, digital music databases, field recordings.

Curation: Segmenting longer audio tracks, normalizing volume, filtering out noise, and labeling based on content.

Video Data:

Sources: Movies, TV shows, online videos, animations, video game footage.

Collection: Video platforms like YouTube, digital video databases.

Curation: Segmenting videos, categorization based on content, and sometimes extracting audio for separate processing.

Scientific and Research Data:

Sources: Research papers, reports, academic publications.

Collection: Online academic databases like JSTOR or Google Scholar, academic institutions' repositories.

Curation: Filtering out irrelevant or outdated papers, categorizing based on subject matter.

Code and Software:

Sources: Open-source repositories, software applications, scripts.

Collection: Platforms like GitHub, GitLab, or proprietary software databases.

Curation: Removing non-functional code, categorizing based on functionality or application.

The collection and curation of such data can sometimes lead to copyright infringements, especially if proper permissions are not obtained. For AI models to be trained legally and ethically:

Permission and Licensing: Always respect licensing agreements. Some datasets may be free for academic use but not commercial use, for instance.

Fair Use Consideration: In some jurisdictions, the use of copyrighted material without permission might be allowed under "fair use" doctrines for specific purposes like research. However, this is a complex area and often requires legal advice.

Public Domain and Open Licenses: Often, researchers and developers prioritize data that's in the public domain or available under open licenses (like Creative Commons licenses) to sidestep these concerns.

The intersection of AI and copyright is a developing area of both technology and law. As AI continues to grow and evolve, so will the methods and best practices for collecting and curating training data.

6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?Developers of AI models acquire materials or datasets for training through various channels. The origin and access largely depend on the nature of the project (commercial,

academic, or hobbyist) and the scale of data required. Here's a breakdown of the avenues:

Public Datasets: Many organizations and institutions release public datasets to further research and development in AI. Examples include:

UCI Machine Learning Repository: A collection of databases, domain theories, and datasets.

Kaggle: Beyond hosting predictive modeling competitions, Kaggle is a collaborative community platform where users can publish datasets, kernels (code), and notebooks, fostering a culture of shared knowledge.

ImageNet: A dataset containing millions of labeled images for object recognition.

Academic Institutions: Universities and research institutions sometimes collect and release datasets specific to their research. These datasets are frequently accessed by students, researchers, and the broader AI community.

Government and NGOs: In the spirit of transparency or to spur innovation in areas of public interest, government agencies and NGOs release datasets. Examples:

U.S. Government's open data: Contains data ranging from agriculture to health.

European Union Open Data Portal: Provides access to diverse data from the institutions and other bodies of the European Union.

Third-party Data Brokers and Private Companies: These businesses specialize in gathering, curating, and selling data. They source data from public records, online activity, and a myriad of other channels.

Self-collection using Web Scraping: Developers sometimes employ tools and scripts to scrape data from websites and online platforms. While this method is prevalent, it comes with legal, ethical, and technical challenges, especially if the data is

copyrighted, scraping goes against a website's terms of service, or the website deploys anti-scraping measures.
Collaborations and Partnerships: Larger organizations, especially tech giants, often form partnerships to access proprietary data. This might be through business collaborations, academic partnerships, or dedicated research initiatives.

APIs: Many platforms, like Twitter and Reddit, offer APIs (Application Programming Interfaces) that let developers collect data in a structured format. Usage often requires an API key, and while there are rate limits and terms of use, certain platforms may impose costs or other restrictions for high-volume or commercial use.

Purchase or Licensing: For niche requirements, companies might opt to directly purchase or license data from its creators or data marketplaces. Such transactions typically involve legal contracts to ensure both parties' interests, especially when handling sensitive or proprietary information.

Crowdsourcing: Platforms such as Amazon Mechanical Turk facilitate data collection by harnessing a large pool of human contributors. It proves invaluable for tasks like image labeling, transcription, and more.

Regarding the extent to which third-party entities first collect training material:

Significant Role of Third Parties: Many AI startups and teams, especially those with limited resources, heavily rely on third-party datasets from academic researchers, public sources, or private entities. These datasets not only offer a foundational starting point but also bring a degree of quality and diversity, critical for enhanced AI model performance.

Customized Data Collection: Although third-party datasets are pivotal, unique AI applications might demand customized data collection. Here, entities might resort to self-collection, forming partnerships, or procuring tailored datasets.

In summary, while a vast reservoir of publicly available data exists, groundbreaking work in AI often necessitates a concoction of public datasets, proprietary data, and fresh data collection, occasionally involving third-party entities. Proper attribution, strict adherence to data privacy regulations globally, and respecting copyright and licensing agreements remain paramount in all these endeavors.

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

The use of copyrighted works as training materials for AI systems is a complex issue, encompassing both ethical and legal dimensions. Here's an exploration of the topic:

Extent of Use of Copyrighted Works as Training Materials:

● Widespread Usage: Many AI models, especially those trained for natural language processing (like ChatGPT) or image recognition, are trained on vast amounts of data. Some of this data may come from copyrighted sources, such as books, articles, websites, photos, and other media.
● Grey Areas: Sometimes, the use of copyrighted works for training AI models lies in a gray area. For instance, while individual data points (like a sentence or an image) might be copyrighted, the patterns the AI learns from aggregating vast amounts of such data may not be. The act of training on copyrighted data does not necessarily mean the output will infringe on that copyright.

Licensing of Copyrighted Works for Training:

● Direct Licensing: Some companies directly license data from copyright holders. This is more common for specific projects where the accuracy and source of the data are paramount. For instance, medical AI systems may be trained on licensed medical imagery.
● Creative Commons and Open Licenses: Many datasets leverage materials under Creative Commons or other open licenses. These licenses often allow uses like training AI models, provided certain conditions (like attribution) are met.

● Fair Use: In some jurisdictions, the use of copyrighted materials for research, including AI training, might be considered "fair use." However, this is a complex legal argument and varies from one jurisdiction to another.

Licensing Models Being Offered and Used:

● Subscription-Based Access: Some data providers offer subscription models, where AI developers can access and use datasets for a recurring fee.
● One-Time Licensing: A more traditional model where datasets are licensed for a one-time fee.
● Freemium Models: Some providers may offer basic datasets for free and charge for more comprehensive or specialized datasets.
● Custom Data Collection: Licensing agreements where the data provider collects and provides data tailored to the AI developer's needs.
● Data Sharing Agreements: Especially common in academia, where institutions might agree to share data under specific conditions, ensuring mutual benefit without direct monetary transactions.
● Public Domain Datasets: Datasets explicitly released in to the public domain, allowing unrestricted use.

It's crucial to note that licensing copyrighted works for AI training is still an evolving field. As AI becomes more pervasive and its societal implications more evident, copyright owners, AI developers, and legal systems are bound to further refine the standards and practices around data licensing.

Lastly, while licensing models provide a legal framework, ethical considerations are also paramount. Ensuring data privacy, respecting the intent of original creators, and recognizing the potential biases in training data are all essential aspects of responsible AI development.

6.3.To what extent is non-copyrighted material (such as public domain works) used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?

Utilization of Non-Copyrighted Material for AI Training:

● High Utility of Public Domain Works: Public domain works are a valuable resource for AI training. Since they're not bound by copyright restrictions, they

provide a rich, hassle-free source of data. Examples include classic literature, historical documents, and older academic papers, which are used extensively, especially in natural language processing models.
● Open Access Repositories: There are numerous datasets in the public domain specifically curated for various AI applications. Datasets like Project Gutenberg (for books) or ImageNet (for images) contain vast amounts of non-copyrighted materials and are foundational for many AI research projects.
● Government and Public Records: Many government publications, datasets, and records are publicly accessible and not copyrighted. They provide a treasure trove of data ranging from census details, weather data, to legal rulings.

Creation and Commissioning of Training Material by AI Developers:

● Custom Datasets for Specific Tasks: AI developers often need specific types of data that aren't readily available. In such cases, they may commission the creation of custom datasets. For instance, a company developing a voice assistant might hire people to record specific phrases to train its model.
● Synthetic Data Generation: With advancements in AI, there's a growing trend of using AI to create synthetic data for training other AI models. GANs (Generative Adversarial Networks) are a prime example, where one AI generates data and another evaluates it, iterating until the generated data is of high quality.
● Augmentation: AI developers use techniques to augment existing data, making it more diverse and extensive. For instance, an image can be rotated, zoomed, or color-adjusted to create variations.
● Collaborative Data Collection: Platforms like Kaggle have hosted challenges where the community is incentivized to collect and share data on specific tasks. Such collaborative efforts can lead to large, diverse datasets created explicitly for AI training.
● Simulations and Virtual Environments: For tasks like training autonomous vehicles or robotic arms, simulations can generate vast amounts of training data. Virtual environments can be controlled and adjusted to produce diverse scenarios.

In conclusion, while copyrighted material poses certain challenges and ethical considerations for AI training, the vast expanse of public domain works and the

ability of AI developers to create or commission specific datasets offer multiple avenues to acquire training data responsibly. As AI continues to evolve, the strategies for data acquisition will likely become more refined, ensuring that models are both effective and ethically trained.

6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.

Retention of Training Materials by AI Developers:

The practice of retaining training materials after the completion of an AI model's training varies among developers and organizations. The decision often hinges on several factors, including the purpose of the model, storage costs, future needs, and regulatory considerations. Here are some reasons and practices associated with the retention of these materials:

- Iterative Refinement:
  - Purpose: Training an AI model is rarely a one-time process. Models may be refined and retrained periodically to improve accuracy or to account for new data.
  - Practice: Developers store the original datasets so they can easily reintroduce them during these iterative training processes without having to reconstruct or reacquire the data.
- Model Verification and Validation:
  - Purpose: Keeping training data can assist in verifying and validating the model's behavior in the future.
  - Practice: By retaining the training set, developers can re-run the model to verify its results at any point, ensuring its outputs remain consistent.
- Regulatory and Compliance Needs:
  - Purpose: Certain industries(e.g., finance, healthcare) have strict regulatory requirements necessitating transparency and traceability in AI decision-making.

◯ Practice: By keeping training datasets, organizations can demonstrate the data sources their models were trained on, which is vital for audits or regulatory checks.

● Backup and Redundancy:
◯ Purpose: Data loss is a concern in the digital age. Retaining training materials ensures that important data is not lost due to technical glitches or unforeseen events.

◯ Practice: Developers typically maintain multiple backup copies of their data in different locations or on cloud platforms.

● Future Research and Development:
◯ Purpose: Training materials can be useful for future projects or research, potentially saving time and resources in data acquisition.

◯ Practice: Organizations archive datasets in data warehouses or data lakes, categorizing them for easy retrieval in future projects.

● License Agreements:
◯ Purpose: Some datasets come with licensing agreements that dictate storage, usage, and retention policies.

◯ Practice: Developers adhere to the terms of these licenses, which may sometimes mandate prolonged retention or, conversely, timely deletion of the data after use.

● Storage Costs and Considerations:
◯ Purpose: While storage costs have been reducing over the years, storing vast amounts of data can still be expensive, especially for longer durations.

◯ Practice: Organizations perform a cost-benefit analysis to decide on the duration of data retention. Less critical datasets might be purged to save on storage costs.

● Ethical and Privacy Concerns:
◯ Purpose: Data privacy laws, like GDPR or CCPA, have strict guidelines about data storage, especially if the data contains personal information.

◯ Practice: Developers anonymize or pseudonymize sensitive data. In some cases, datasets are deleted post-training to maintain user privacy and to comply with legal mandates.

In essence, the retention of training materials by AI developers is a nuanced process influenced by a blend of technical, regulatory, economic, and ethical considerations. As AI adoption grows, so too will the importance of robust data management practices.

The training of AI models, particularly in the context of deep learning (a subfield of machine learning), is a process that involves refining model parameters based on data. Here's a brief overview of my understanding:

● Dataset Collection and Preparation:
◯ Before training starts, one must acquire and preprocess a dataset. This dataset usually consists of input-output pairs. In the context of supervised learning (the most common approach), these pairs represent examples for the AI to learn from.
◯ Data preprocessing might include normalization (scaling inputs), augmentation (increasing dataset size by making slight modifications to existing data), and encoding (e.g., turning categorical variables into a format suitable for machine learning).
● Model Architecture Selection:
◯ Depending on the problem at hand, a specific model architecture is chosen. For image recognition, Convolutional Neural Networks (CNNs) might be selected, while for sequence data like text, Recurrent Neural Networks (RNNs) or Transformers might be more appropriate.
● Training:
◯ During training, the model is exposed to the data multiple times. Each exposure is termed an "epoch".

◯ For each input in the training data, the model produces an output. This output is compared to the desired output, and the difference is termed the "loss" or "error".

◯ The model's parameters are then adjusted to minimize this loss using optimization algorithms, with gradient descent being the most popular.

● Validation and Overfitting:

◯ While training, it's essential to monitor the model's performance on unseen data to prevent overfitting (where the model performs exceedingly well on training data but poorly on new data). A separate dataset, called the validation set, is used for this.

◯ Techniques like dropout, early stopping, and regularization can help mitigate overfitting.

● Testing:

◯ After training, the model's performance is evaluated on a test dataset that it has never seen before. This gives an indication of its real-world performance.

● Hyperparameter Tuning:

◯ AI models come with hyper-parameters that aren't learned from the training process but are set beforehand. Examples include learning rate, batch size, and number of layers.

◯ Hyperparameter tuning involves experimenting with different values to find the optimal settings for a particular problem.

● Deployment:

◯ Once satisfied with the model's performance, it can be deployed in a real-world setting, be it a software application, a web service, or embedded in devices.

● Continuous Learning:

◯ In some applications, it's beneficial for the model to continue learning as new data becomes available. This ongoing training can help the AI adapt to changing conditions.

7.1. How are training materials used and/or reproduced when training an AI model? Please include your understanding of the nature and duration of any reproduction of works that occur during the training process, as well as your views on the extent to which these activities implicate the exclusive rights of copyright owners.

The process of using training materials to train an AI model involves several steps, many of which necessitate the reproduction of copyrighted works in various ways. Here's a comprehensive breakdown:

- Data Ingestion:
  - Training an AI model typically begins with ingesting or loading the training materials into the system. This often requires the digital reproduction of copyrighted works as they are transferred from their source locations (e.g., databases, files) to the memory of the machine performing the training.
- Preprocessing and Augmentation:
  - Before feeding data into the model, it often undergoes preprocessing to be rendered in a format suitable for the model. For instance, images might be resized or normalized, and text might be tokenized or encoded.
  - Data augmentation, especially in the domain of image and audio data, involves creating modified versions of the original data to enhance the dataset's diversity. This can include rotating images, adding noise to audio, or using synonyms in text. Such modifications may result in the creation of derivative works based on the original copyrighted content.
- Batch Processing:
  - During training, data is often processed in batches, meaning small subsets of the entire dataset are loaded into the model's memory sequentially. This action involves temporary reproductions of the data in system memory.

◯ This batch-wise reproduction occurs repeatedly throughout the training process, every time the model goes through the data.

● Model Checkpointing:

◯ At regular intervals, the current state of the model (its weights and architecture) might be saved for later use or recovery. While the model itself is a set of numerical values and doesn't reproduce the training content, its state is informed by the training data it has seen. Some argue this "knowledge" could be seen as a reflection or transformation of the copyrighted material.

● Duration of Reproduction:

◯ Reproductions of copyrighted materials in system memory are transient. They exist temporarily during the training process and are overwritten or deleted afterward.

◯ However, the final trained model, the datasets, and any augmented or preprocessed data may be stored indefinitely, depending on the requirements and practices of the organization or individual conducting the training.

● Implication for Copyright Owners:

◯ There productions occurring during AI training, particularly transient ones in system memory, present a nuanced challenge for copyright law. While they are technical reproductions, they might not be directly consumable or recognizable in their reproduced form.

◯ The more pressing concern for many copyright holders is the potential for AI models to generate outputs that closely mirror or recreate copyrighted content. This raises questions about whether such outputs constitute infringement.

◯ Further more, if AI is trained on copyrighted materials without proper licensing or permission, even if the direct content is not reproduced in

outputs, there's an argument that the "essence" or "knowledge" of the copyrighted work is captured within the AI.

The use of copyrighted materials in AI training undeniably intersects with the exclusive rights of copyright owners. The crux of the matter lies in determining which reproductions are substantial and actionable and how the evolving capabilities of AI fit within the traditional framework of copyright law. This is a dynamic area of discourse, with technological advances often outpacing legal clarifications.

7.2. How are inferences gained from the training process stored or represented within an AI model?

The inferences gained from the training process are encapsulated within an AI model's structure, specifically in the model's weights and biases. Let's break down how this is achieved:

- Model Architecture:
  - AI models, especially deep neural networks, consist of layers of interconnected nodes (or neurons). The strength and nature of connections between these nodes are represented by values known as weights.
- Training Process:
  - During training, the model adjusts these weights based on the differences between its predictions and the actual results from the training data. This adjustment process, typically facilitated by algorithms like gradient descent, iteratively refines the weights to minimize the model's error.
- Weights and Biases:
  - At the conclusion of the training process, the adjusted weights and biases capture the learned patterns and relationships from the training data. These weights and biases, which are merely numerical values, effectively store the inferences the model has gleaned from the data.

● Model Checkpoints and Serialization:
○ These learned weights and biases are stored in model checkpoint files or are serialized into specific formats suitable for deployment. These files act as a snapshot of the model's knowledge at a particular point in its training and can be loaded later to resume training or for deployment in real-world applications.

● Inferences and Outputs:
○ When new, unseen data is input in to the trained model, it processes this data using its stored weights and biases, resulting in an output or prediction. This output is the manifestation of the inferences and knowledge the model gained during training.

● Abstraction and Non-explicitness:
○ It's crucial to note that the knowledge within an AI model is abstract and non-explicit. Unlike a database that stores explicit information, an AI model's weights don't directly represent specific pieces of the training data. Instead, they capture generalized patterns, allowing the model to perform tasks like classification, prediction, or generation.

In essence, the inferences gained from the training process are stored as numerical weights and biases within the model's architecture. These values determine how the model responds to new inputs, allowing it to make predictions or generate outputs based on its training. The precise and detailed nature of these weights makes it challenging, if not impossible, to reverse-engineer or extract the original training data from a trained model.

7.3. Is it possible for an AI model to “unlearn” inferences it gained from training on a particular piece of training material? If so, is it economically feasible? In addition to retraining a model, are there other ways to “unlearn” inferences from training?

Yes, it's possible for an AI model to "unlearn" certain inferences, but the methods and feasibility can vary.

● Retraining:
○ The most straight forward way to make an AI model "unlearn" specific inferences is to retrain it without the inclusion of the contentious data. By exposing the model to a modified dataset and reiterating the training process, the model's weights and biases adjust, and the previous inferences may diminish or disappear.

● Fine-tuning:
○ Instead of retraining the entire model, one can also fine-tune the model using new data that counteracts the undesired inferences. This approach is more efficient than full retraining but may not guarantee complete "unlearning" of undesired patterns.

● Regularization Techniques:
○ Techniques like dropout, weight decay, and early stopping can be employed during training to prevent the model from overly relying on specific patterns, potentially making it easier for the model to "unlearn" certain inferences.

● Elastic Weight Consolidation(EWC):
○ EWC is a method designed to help neural networks overcome catastrophic forgetting. While its primary use is to retain learned knowledge when training on new tasks, it can be adapted to encourage the model to forget specific inferences by emphasizing other tasks.

Economic Feasibility:

The economic feasibility of "unlearning" largely depends on several factors:

● Model Size and Complexity: Larger, more complex models like GPT-3 or BERT demand more computational resources, making the retraining or fine-tuning process more expensive.
● Dataset Size: Retraining on a massive dataset requires more computational time and resources.
● Urgency: If there's an urgent need to "unlearn" specific information, more resources might be dedicated, raising costs.

Alternative Methods:

● Data Augmentation: Instead of removing the specific datapoint, one can augment the dataset with additional data that counteracts or balances the undesired inferences.
● Negative Reinforcement: In reinforcement learning scenarios, negative reinforcement can be applied to discourage certain behaviors or inferences.
● Knowledge Distillation: This involves training a secondary model (a student) using the outputs of the primary model (the teacher) as guidance. By selectively curating the teacher's outputs or the training dataset, one might push the student model to "unlearn" certain aspects of the teacher model.

In conclusion, while it's technically possible for AI models to "unlearn" specific inferences, the process is not always straightforward or economically feasible. The approach chosen would be contingent on the nature of the undesired inference, the model's architecture, and the available resources.

7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?

Identifying whether an AI model was trained on a specific piece of training material without having access to the underlying dataset is a challenging task. In general, this would be quite difficult for a variety of reasons:

● Abstraction of Information: AI models, especially deep learning models, abstract information in the training process. They don't "remember" specific data points but rather learn patterns, relationships, and structures from the data. Consequently, the exact input data is abstracted to high-level features, making it nearly impossible to reverse-engineer the model to identify specific pieces of training material.
● Volume and Diversity of Data: Given the vast amount of data typically used to train sophisticated models, even if a model produces an output similar to a known piece of training material, it's hard to ascertain if that exact material was part of the training set or if the model is generalizing from other similar data points.

● Overfitting vs. Generalization: If a model is overfitting, it might closely memorize training data, making it potentially more susceptible to revealing specifics about that data. However, well-trained models are designed to generalize and not memorize, so they're less likely to exhibit outputs that can be directly linked to specific training examples.

However, there are some emerging areas of research and methodologies that can provide insights:

● Model Inversion Attacks: These are attempts to recreate a piece of training data given a trained model. While it's challenging to retrieve an exact piece of data, it might be possible to obtain something similar, especially if the model is overfitting.
● Membership Inference Attacks: This type of attack aims to determine whether a particular data point was part of the training set. While it doesn't recreate the actual data, it can give an indication of its presence.
● Extraction and Analysis of Model Weights: In some cases, if the model weights and structures are accessible, researchers might be able to glean information about the nature of the training data, though not specifics about individual data points.
● Analyzing Model Responses: By interacting with the model (e.g., asking specific questions to a language model), one might be able to deduce some information about its training data based on the patterns in its responses. However, this method is indirect and speculative.

In summary, while there are methods to attempt to deduce or infer specifics about a model's training data, reliably and precisely identifying a particular piece of training material used in the model's training, absent access to the dataset, is an uphill challenge. However, as research in this area progresses, the capabilities to extract or infer such information might evolve.

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.

● The doctrine of fair use, codified at 17 U.S.C. § 107, allows for the limited use of copyrighted material without requiring permission from the rights holders.

This essential principle in U.S. copyright law seeks to balance the interests of copyright holders not only with the public's interest in the broader distribution and use of creative works but also with other societal values such as critique, comment, news reporting, teaching, scholarship, and research.

● In the context of AI and machine learning, the unauthorized use of copyrighted works to train models ventures into uncharted legal territory. To determine fair use, courts typically consider four primary factors:

● Purpose and character of the use: Transformative uses, which add something new or alter the original work's character, are more likely to be deemed fair use. The commercial aspect of the use is also pertinent, with non-commercial uses being more favored. AI training can be perceived as transformative since it processes the work to extract patterns and insights rather than replicating the work. Nonetheless, if the AI model's application is commercial, this might weigh against a finding of fair use.

● Nature of the copyrighted work: Works that are factual or non-fictional are more susceptible to fair use compared to highly creative endeavors like fiction, music, or art. Thus, training AI models on factual databases or compilations might be more defensible than using sources like novels or songs.

● Amount and substantiality of the portion used: Using smaller or less significant portions of a copyrighted work can favor a finding of fair use. However, AI training often involves large datasets, potentially encompassing significant portions of copyrighted materials.

● Effect on the market or potential market for the copyrighted work: This factor, often deemed crucial by courts, considers if the use harms the copyright holder's income or potential market. If AI training on copyrighted content doesn't undermine demand for the original, this factor might support a fair use finding.

Relevant Case Law:

● Authors Guild, Inc. v. Google, Inc.: The Second Circuit determined that Google Books' digitization of copyrighted works for search purposes was a transformative use. They emphasized that Google Books doesn't offer the entire copyrighted books to the public but instead provides information about them.

● Kelly v. Arriba Soft Corp.: The Ninth Circuit found that using copyrighted images as search engine thumbnails was transformative. The images served to help users identify and locate images, not to communicate the original artistic expression.

8.1. In light of the Supreme Court's recent decisions in *Google* v. *Oracle America* and *Andy Warhol Foundation* v. *Goldsmith,* how should the "purpose and character" of the use of copyrighted works to train an AI model be evaluated? What is the relevant use to be analyzed? Do different stages of training, such as pre-training and fine-tuning, raise different considerations under the first fair use factor?

The Supreme Court's decisions in Google v. Oracle America and Andy Warhol Foundation v. Goldsmith can offer some insights into how courts might evaluate the "purpose and character" of using copyrighted works to train an AI model.

Google v. Oracle America: In this landmark decision, the Supreme Court found that Google's copying of a portion of the Java SE Application Programming Interface (API) to create a new Android mobile operating system was a fair use under copyrightlaw.TheCourtemphasizedthetransformativenatureofGoogle'suse, noting that Google used the copied lines for a new and different purpose — to create a new platform that could be readily used by programmers.

Implications for AI Training: If an AI model's training uses copyrighted material for a similarly transformative purpose — i.e., not to replicate the work but to extract patterns, learn from them, and create something new — this could weigh in favor of fair use. However, the sheer volume of data needed for AI training might differ from the more limited use of Java API lines by Google, and this could affect the analysis.

Andy Warhol Foundation v. Goldsmith: This case centered on Warhol's series of silkscreen and pencil illustrations based on a photograph of the musician Prince, taken by Goldsmith. The Second Circuit held that Warhol's works were not transformative and thus not protected by fair use. The court emphasized that just because a work adds something new does not make it transformative in a way that could justify a fair use defense.

Implications for AI Training: AI's training process might add something new or change the character of the copyrighted material. Still, it's essential to consider whether this change is sufficiently transformative in the eyes of the law. Simply processing and storing data in a new format or context might not meet the threshold set by the Goldsmith case.

Pre-training vs. Fine-tuning: Different stages of AI training could indeed raise varying considerations under the first fair use factor.

● Pre-training: This involves training models on vast amounts of data to learn general patterns, relationships, and structures. If copyrighted materials are extensively used in this stage, the sheer volume and potential lack of transformation could weigh against a finding of fair use.
● Fine-tuning: In this stage, models are further trained on more specific data, often tailored to particular tasks. If copyrighted materials are used more selectively and for specialized transformative purposes during fine-tuning, it might be easier to argue for fair use.

In light of the cited decisions, the “purpose and character” analysis for AI training should thoroughly evaluate the transformative nature of the use, the newness of the created output, and the specific context and manner in which copyrighted materials are employed. The exact details of AI training, including the distinct stages and their requirements, would play a crucial role in this assessment.

8.2. How should the analysis apply to entities that collect and distribute copyrighted material for training but may not themselves engage in the training?

Entities that collect and distribute copyrighted material for AI training, without directly engaging in the training, play a crucial intermediary role. Their actions can have significant implications for copyright law, especially in the context of the fair use analysis. Here's how the evaluation might be approached:

Nature of Distribution: The first step is to understand the nature of the distribution. If these entities are actively encouraging the use of copyrighted materials for AI training without acquiring the necessary rights or licenses, they might be indirectly facilitating copyright infringements.

Commercial vs. Non-commercial Intent: The “purpose and character” of the distribution are central to the fair use analysis. If the collection and distribution are done with a commercial intent — for profit, for instance — it could weigh against a finding of fair use. In contrast, a non-profit organization distributing data for academic research might be seen more favorably under the fair use doctrine.

Transformative Use: Even if the entity is not directly training the AI, the potential transformative use of the material by the end-users should be considered. If the entity provides tools, guidelines, or context that facilitate a transformative use of the material, it may bolster a fair use argument. Conversely, if they distribute copyrighted material with the suggestion or implication of direct replication in AI systems, it could lean against fair use.

Accountability and Due Diligence: Entities should exhibit due diligence in sourcing and distributing copyrighted materials. Proper attribution, where possible, and transparent communication about the copyright status of distributed materials can play in their favor. Ignorance or blatant disregard for copyright, on the other hand, can be detrimental in any legal evaluation.

Licensing and Permissions: If the entity has sought permissions or licenses to collect and distribute copyrighted material, it will be in a much stronger position. Even if they aren’t directly engaging in AI training, demonstrating a proactive approach to copyright compliance can be a significant factor.

Indirect Liability: Even if they don't directly infringe copyrights, these entities can be seen as contributors or facilitators to potential infringements, especially if they are aware of, or have a financial interest in, the infringing activities.

In conclusion, while these entities may not engage in AI training themselves, their activities in collecting and distributing copyrighted material can have significant copyright implications. The evaluation would revolve around their intent, the nature of their distribution, their approach to copyright compliance, and the potential transformative use of the materials they distribute.

8.3. The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes. How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? Does it make

a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems?

The transition from noncommercial or research purposes to commercial utilization of AI models trained on copyrighted materials introduces complexities into the fair use analysis. Here's a breakdown of the factors to consider:

1. Original Intent vs. Subsequent Use: If copyrighted materials were initially used for research or noncommercial purposes, but the resulting AI model or dataset is later used commercially, it could impact the fair use justification. Even if the initial intent was noncommercial, the subsequent commercial use could weigh against a finding of fair use, especially if the commercial exploitation is significant.

2. Degree of Transformation: A key tenet of fair use is the concept of transformation. If the copyrighted material is used as a foundation for creating something distinctly new or has undergone significant transformation during the AI training, then it might strengthen the fair use argument, even in a commercial context.

3. Nature of the Copyrighted Work: Using more factual or non-fictional copyrighted materials might be looked upon more favorably in a fair use analysis than using highly creative works. The rationale is that factual or informational content serves a broader societal interest.

4. Amount and Substantiality: Even if the training was initially for research, if the entirety of a copyrighted work (or its most crucial parts) is used, it could weigh against fair use, especially if it's now being used for commercial gain.

5. Impact on the Market: One of the essential considerations is whether the commercial use of the AI model or dataset affects the potential market for, or the value of, the copyrighted work. If it does, it could be detrimental to a fair use claim.

6. Funding Source: The source of funding could influence the fair use analysis. If a noncommercial or research project is funded by for-profit developers of AI systems, it could be seen as an indirect commercial endeavor, especially if the end goal is to integrate the research outcomes into commercially available products or services.

In summary, transitioning from noncommercial to commercial use doesn't automatically negate a fair use argument, but it does introduce additional scrutiny. The nature of the copyrighted work, the degree of transformation, the potential

market impact, and the nuances of funding all play crucial roles in determining how fair use might be applied in such contexts.

8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how?

Quantity of Training Materials Used in AI: Generative AI models, especially those in deep learning realms such as GANs (Generative Adversarial Networks), often require vast amounts of data for effective training. The rationale behind using large datasets is to capture as much variability and richness in the data as possible, ensuring that the model generalizes well to novel scenarios and produces diverse, high-quality outputs. For some sophisticated models, training data can span millions to billions of data points.

Implications for Fair Use:

- Amount and Substantiality: One of the core factors in the fair use analysis is the amount and substantiality of the portion used in relation to the copyrighted work as a whole. Using a significant portion of a copyrighted work might weigh against a fair use claim, especially if that portion embodies the heart of the content.
- Proportionality and Purpose: While the sheer volume might seem concerning, it's essential to consider proportionality. If a vast dataset consists of millions of copyrighted works, each individual work's impact might be minor in the broader context of the entire dataset. Moreover, the purpose of using such a large volume is not to replicate or replace individual copyrighted content but to understand patterns and structures within the data.
- Transformation: Given the nature of generative AI models, the input(training data) undergoes a transformation, resulting in outputs that are often novel and different from the training data. If the model's output is significantly transformed, it could bolster the argument for fair use.
- Market Effect: The last factor in the fair use analysis considers the effect of the use upon the potential market for or value of the copyrighted work. If the vast volume of data used doesn't negatively impact the market value of individual copyrighted pieces within the dataset, it could lean towards fair use.

Conclusion: While the volume of training material is undoubtedly a point of consideration, the fair use analysis is multifaceted. The sheer quantity of training data in AI models can be a double-edged sword – while it provides the model with comprehensive knowledge, it simultaneously poses challenges for copyright considerations. It's crucial to evaluate the quantity in the broader context of purpose, transformation, and market impact.

8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

Understanding the Fourth Factor: The fourth factor in the fair use analysis assesses the effect of the use on the potential market for or value of the copyrighted work. This factor often holds significant weight because, at its core, copyright law aims to provide creators with economic incentives to continue producing work.

Measuring the Effect in the AI Context:

● Direct Competition with Original Work: A primary way to measure market impact is by evaluating whether the outputs from the AI system directly compete with the original copyrighted material. If an AI system trained on a copyrighted novel starts producing similar novels, it directly affects the market for the original work.
● Impacting the Body of Works: Sometimes, the use might not affect a single work but might harm the broader portfolio of an author or creator. For instance, if a music-generating AI is trained extensively on a particular artist's songs and begins to produce songs with a remarkably similar style, it could diminish the value of that artist's entire body of work.
● Effect on the General Class of Works: An AI system's outputs might not compete directly with specific copyrighted works or even an author's broader portfolio but might saturate the market for a general class of works. For example, if an AI trained on various romantic novels starts producing a

massive volume of romantic stories, it could dilute the market for all romantic novelists.
● Secondary and Tertiary Market Effects: Beyond primary sales, copyrighted works often have secondary markets – such as adaptations, merchandising, or derivative works. If AI outputs begin to substitute or diminish the value in these secondary markets, it's an essential consideration for the fourth factor.
● Temporal Considerations: The timing of market impact is also crucial. While immediate effects on the market are evident, long-term repercussions - like the potential stifling of innovation due to diminished economic incentives for creators - should also be factored into the analysis.

Concluding Thoughts: When assessing the market effect in the context of AI training, it's not just about direct competition with individual copyrighted works. The broader implications for authors' portfolios, the general class of works, and secondary markets should all be part of the evaluation. As AI systems become more sophisticated and their outputs more varied, understanding and assessing these market impacts will be a complex yet vital component of the fair use discussion.

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?

The Foundations of Consent in Copyright: Consent is a cornerstone of copyright law, offering creators control over the use of their intellectual creations. With the rise of AI and its voracious appetite for data, how this consent is structured—either as an opt-in or opt-out mechanism—brings up compelling arguments on both sides.

Opt-In Approach: Ensuring Active Consent:

● Respect for Copyright Holders: An opt-in approach respects the rights of copyright holders by ensuring their works aren't used without explicit permission. This method aligns closely with traditional copyright principles where any use requires prior consent.
● Clarity and Predictability: With opt-in, AI developers would have clear guidelines—only use works for which they have obtained explicit permission. This reduces the gray areas in AI training processes.

● Limitations: While it respects copyright principles, an opt-in approach may drastically limit the volume and variety of data available for AI training, potentially stunting AI innovation and efficiency.

Opt-Out Approach: Balancing Progress and Rights:

● Promotion of AI Innovation: By defaulting to an opt-out mechanism, there would be a larger pool of materials available for AI training, fostering innovation and the development of more sophisticated models.
● Responsibility on Copyright Holders: Placing the onus on creators to object or opt-out could be seen as burdensome, especially for those who might not be aware of their works being used.
● Potential for Abuse: Without active consent, there's a risk that copyrighted works might be used in ways that original creators did not anticipate or might not approve of if they knew.

Balancing the Scales:

● Hybrid Models: One potential solution is a hybrid model, combining elements of both opt-in and opt-out. For instance, certain categories of works (like academic research papers) could be opt-out, while others (like personal photographs) could be opt-in.
● Centralized Databases: A central database could be developed, where creators can easily register their preference. This would reduce the burden on individual copyright holders and provide AI developers with a clear list of what's available.

Conclusion: The decision between opt-in and opt-out isn't black and white. While an opt-in approach aligns closely with traditional copyright principles, the exponential growth and potential of AI suggest that some adaptation of these principles might be necessary. The challenge lies in finding a balance that respects the rights of creators while also fostering innovation in the AI sphere. As AI continues to permeate our daily lives, this discussion will only become more crucial, and it's essential to engage all stakeholders in the dialogue.

9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses?

The Core of Copyright Consent: At the heart of copyright law is the concept of consent: the right of creators to have a say over how their creations are utilized. However, the emergence of AI, especially in the domain of training models, poses the question of when consent should be mandated. Should it be for all uses, or limited only to commercial exploits?

Consent for All Uses: Prioritizing the Rights of Creators:

● Upholding Traditional Copyright Values: Mandating consent for all uses upholds the sanctity of copyright principles, ensuring that creators maintain control over their work irrespective of the end purpose.
● Avoiding Ambiguity: If consent is required for all uses, it removes any ambiguity, making it easier for both AI developers and copyright holders to understand and follow the rules.
● Potential Stifling of Innovation: While this approach is protective of creators' rights, it may deter academic research or not-for-profit ventures from using copyrighted materials, even when their intent is non-commercial and perhaps beneficial for society.

Consent Only for Commercial Uses: Striking a Balance:

● Encouraging Research and Development: Exempting non-commercial uses from mandatory consent can boost academic research, educational endeavors, and other not-for-profit activities. This distinction acknowledges the difference between profiting from someone's work and using it for broader societal benefits.
● Complexity in Definition: What precisely constitutes a "commercial" use? The line between commercial and non-commercial can sometimes be blurry, especially in cases where research has potential future commercial applications.
● Potential Exploitation: There's a risk that certain entities might disguise their commercial intentions as non-commercial to bypass obtaining consent, leading to potential misuse.

Considerations for a Nuanced Approach:

● Evolving Nature of Projects: Some projects might start as non-commercial (e.g., academic research) but evolve into commercial applications. How should consent be managed in such cases?
● Layered Consent Mechanisms: One could consider a mechanism where works are freely available for non-commercial uses, but any shift towards a commercial application triggers the requirement for consent.
● Awareness and Education: Educating copyright holders about the nuances of AI and its potential applications can help them make informed decisions about granting permissions.

Conclusion: The decision between requiring consent for all uses versus only commercial ones is intricate. While prioritizing the rights of creators is essential, it's equally important to foster an environment conducive to innovation and societal advancement. The optimal solution may lie in crafting a nuanced approach, informed by continuous dialogue among creators, AI developers, and other stakeholders.

9.2. If an "opt out" approach were adopted, how would that process work for a copyright owner who objected to the use of their works for training? Are there technical tools that might facilitate this process, such as a technical flag or metadata indicating that an automated service should not collect and store a work for AI training uses?

Understanding "Opt-Out": An "opt-out" approach inherently assumes a default permission for copyrighted works to be used in AI training. Under this system, the onus is on the copyright owner to expressly indicate their objection to their work being used in such a manner.

How Could an "Opt-Out" Approach Work?:

● Centralized Repository:
○ A centralized platform or database could be established where copyright owners register their works.
○ Each registered work could have a default setting allowing its use in AI training.
○ Owners can toggle or select an "opt-out" option for works they don't want to be used.

● Technical Tools and Metadata:
○ Copyrighted works, especially digital ones, can carry metadata. This metadata can be structured to include information regarding the consent status for AI training use.
○ Automated systems that scrape or collect data for AI training can be programmed to read this metadata. If the metadata indicates an "opt-out" status, the system would skip or exclude that particular work.
○ The use of blockchain technology could further ensure the authenticity and non-tamperable nature of such metadata.
● AI-Driven Monitoring:
○ Paradoxically, AI could be part of the solution. AI systems can be trained to monitor and detect unauthorized use of copyrighted materials in AI training datasets across the web.
○ Once identified, these systems can notify copyright owners of potential infringements.

● Clear Communication Channels:
○ Platforms where copyrighted works are commonly sourced should provide clear mechanisms for owners to communicate their opt-out decisions. For instance, a photo-sharing platform might have a checkbox indicating "Do not use for AI training."
● Regular Audits and Transparency Reports:
○ Companies involved in AI training could perform regular audits to ensure compliance with the opt-out wishes of copyright owners.
○ Transparency reports detailing sourcing practices for AI training datasets can bolster trust among creators and the public.

Challenges and Considerations:

● Scale: Given the sheer volume of copyrighted works and the rapid pace of content creation, maintaining an up-to-date opt-out system could be daunting.

● Awareness: For an opt-out system to be effective, copyright owners need to be adequately informed about their rights and the implications of AI training on their works.
● Enforcement: Mechanisms need to be in place to handle cases where copyrighted works are used against the wishes of their creators.

Conclusion: While the “opt-out” approach can streamline the sourcing process for AI training materials, it comes with its own set of challenges. Any implementation should prioritize the rights and wishes of copyright owners while leveraging technology to ensure a smooth and respectful handling of copyrighted content.

9.3. What legal, technical, or practical obstacles are there to establishing or using such a process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners?

Legal Challenges:

● Ambiguous Copyright Status:
○ Many works online might have an unclear copyright status. Determining the ownership and consent for such materials can be legally challenging.

● Jurisdictional Differences:
○ Copyright laws vary from country to country. A process that's compliant in one jurisdiction might not be in another. This can be particularly challenging for AI models trained on global data.

● Enforcement:
○ Once an opt-out system is in place, enforcing compliance and penalizing breaches, especially across international borders, can be legally complex.

Technical Challenges:

● Metadata Handling:

◯ While metadata can indicate copyright preferences, not all platforms or file formats support rich metadata. Also, metadata can be stripped, altered, or overlooked during data transfers.

● Scale:
◯ The sheer volume of data used in training AI models poses a technical challenge. Processing such large datasets to filter out opted-out content in real-time can be computationally intensive.

● Dynamic Web Content:
◯ Web content is dynamic, with works being updated, deleted, or moved. Keeping track of the constantly changing status of copyright preferences can be a technical ordeal.

● False Positives/Negatives by AI Monitoring:
◯ AI-driven solutions to monitor unauthorized use can produce false positives or miss infringements altogether.

Practical Challenges:

● Awareness Among Copyright Owners:
◯ Many content creators might not be aware of the implications of AI training or even the existence of opt-out mechanisms.

● Burden on Content Creators:
◯ Requiring individual creators to register and continually update their preferences might be cumbersome and discourage participation.

● Feasibility of Individual Consents:
◯ Given the vast amount of data used in training AI models, seeking individualized consent from each copyright owner might be impractical, especially for large datasets.

● Database Management:
◯ Maintaining a centralized repository of copyright preferences requires robust database management, regular updates, and potentially significant financial resources.

● Potential Stifling of Innovation:
○ Overly stringent requirements might discourage AI research, especially among smaller entities or individual developers without the resources to navigate complex consent mechanisms.

Is It Feasible to Get Consent in Advance?

Given the vast scale at which AI models operate and the sheer amount of data they require, obtaining individual consents in advance is logistically challenging. For commonly used, vast datasets, it might be nearly impossible to seek and verify consent for every individual piece of content. Instead, a more general system, perhaps based on the type or category of data, might be more feasible.

Conclusion:

While the ethical foundation of seeking consent is solid, the practicalities of doing so in the AI realm are fraught with challenges. Balancing respect for copyright ownership with the practicalities of AI research requires a nuanced approach, combining legal, technical, and ethical considerations.

9.4. If an objection is not honored, what remedies should be available? Are existing remedies for infringement appropriate or should there be a separate cause of action?

When an objection is not honored, determining the appropriate remedies can be a complex issue due to the unique nature of AI training and its implications on copyrighted works. Let's evaluate potential remedies and the appropriateness of existing solutions:

Existing Remedies for Copyright Infringement:

● Statutory Damages:
○ Present copyright law provides for statutory damages for each work infringed. This can act as a deterrent for entities that might otherwise use copyrighted content without permission.

● Injunctions:

◯ Copyright owners can seek to stop the offending AI entity from further use of their copyrighted material through injunctions.

● Attorney's Fees and Costs:
◯ In some cases, the prevailing party in a copyright law suit maybe entitled to recover attorney's fees and costs, which can further act as a deterrent.

● Impounding and Disposition:
◯ Existing laws often allow for the impounding and, in some cases, destruction or other disposition of infringing articles.

Potential New Remedies:

● AI-Specific Statutory Damages:
◯ Given the distinct nature of AI infringement (where direct economic harm may be hard to quantify), it might be apt to establish a separate statutory damage metric specifically for AI-related infringements.

● Mandatory "Unlearning":
◯ Require the AI model to be retrained without the copyrighted material in question. This could be a significant deterrent as retraining can be resource-intensive.

● Transparency Reports:
◯ Infringing parties can be mandated to provide detailed reports on the use of copyrighted content, which could act both as a remedy and a preventive measure.

● Alternative Dispute Resolution (ADR):
◯ Given the technical nuances of AI, a specialized ADR mechanism focusing on AI-related disputes might be more efficient than traditional legal avenues.

● Licensing Penalties:

◯ In cases where copyrighted content is used without honoring objections, infringers could be required to pay licensing fees retrospectively, possibly with additional penalties.

Appropriateness of a Separate Cause of Action:

Given the unique nature of AI and its potential impact on copyrighted works, a strong argument can be made for a separate cause of action specifically tailored to AI-related infringements. This cause could:

● Acknowledge the technical intricacies of AI training.
● Recognize the distinct ways AI might "infringe" without traditional copying.
● Ensure remedies are appropriate and effective for the digital age.

Conclusion:

While existing remedies provide a foundational framework, the evolving nature of AI might necessitate the introduction of new, more tailored remedies. A dedicated cause of action for AI-related copyright infringements could offer a balanced approach, ensuring the rights of copyright owners are protected while also fostering innovation in the rapidly expanding field of artificial intelligence.

9.5. In cases where the human creator does not own the copyright—for example, because they have assigned it or because the work was made for hire—should they have a right to object to an AI model being trained on their work? If so, how would such a system work?

The idea that a human creator, distinct from the copyright owner, might have a say in how their work is used in AI training presents an interesting and complex issue. This situation requires a consideration of moral rights, personal association, and the mechanics of intellectual property law.

Moral Rights Perspective:

In some jurisdictions, artists and authors have what are known as "moral rights" in their creations. These rights can include the right to attribution (to be recognized as the author) and the right to integrity (to object to derogatory treatments of their works which could be prejudicial to their honor or reputation).

● Application to AI: If a human creator feels that training an AI model on their work would somehow harm the integrity of that work, then they might seek to object on the grounds of their moral rights.
● Limitations: Not all jurisdictions recognize moral rights, and where they are recognized, they may not be assignable or waivable. The U.S., for example, has a more limited view of moral rights than some European countries.

Practical Considerations:

● Notification: One of the primary challenges would be notification. If a human creator has the right to object, they would need to be informed whenever their work is being used for AI training.
● Implementation: To make this right actionable, there might need to be a system in place akin to the "opt-out" system for copyright owners. An accessible registry where creators can register objections, possibly linked to works by metadata or unique identifiers, could be one approach.
● Jurisdictional Challenges: Given the global nature of AI training and data sources, implementing and enforcing such a right could be challenging across different jurisdictions.

Balancing Interests:

Providing human creators a separate right to object would need to carefully balance multiple interests:

● AI Developers: They would need clarity on the legal landscape and whose permissions they need to obtain or objections they might face.
● Copyright Owners: These entities or individuals would have concerns about another layer of rights, which might complicate the usage landscape.
● Human Creators: These individuals would value having a say in the legacy and ongoing use of their creative works, especially in novel contexts like AI training.

Conclusion:

If such a system were to be implemented, it would represent a significant departure from traditional intellectual property models, injecting more personal and moral considerations into what is typically a more transactional domain. While it would undoubtedly add complexity, it could also be seen as an acknowledgment of the

deep personal connection many creators feel to their works, even when they no longer own the copyright. The challenge would be in crafting a system that is both respectful of these personal connections and practical in its application.

10. If copyright owners' consent is required to train generative AI models, how can or should licenses be obtained?

Acquiring consent from copyright owners is a critical step in ensuring that the intellectual property rights of creators are respected. The AI domain, with its unique challenges and vast scale, requires innovative approaches to licensing. Here's a breakdown of some potential mechanisms:

1. Traditional Licensing:

● Direct Licensing: AI developers can directly negotiate with copyright owners for licenses. This model is effective for large-scale projects where specific datasets are required.
● Collective Licensing: In some sectors, collective management organizations (CMOs) can grant licenses on behalf of copyright owners. CMOs can simplify the process for AI developers, offering bulk licensing for a range of works.

2. Compulsory Licensing: Some jurisdictions might consider introducing a compulsory licensing system where AI developers can use copyrighted works for training purposes by paying a predetermined fee. While this method guarantees access, it could be contentious if copyright owners feel the fees are not representative of the work's value.

3. Statutory Licensing: Countries can introduce statutory licenses for AI training, similar to how some nations have statutory licenses for broadcasting or cover songs. This would involve AI developers paying a fixed fee for using copyrighted content, potentially distributed among copyright owners through CMOs.

4. Open Licensing:

● Creative Commons Licenses: These licenses, especially ones that allow for modifications (like CC-BY or CC-BY-SA), can be instrumental. AI developers could use works under these licenses for training, provided they adhere to the license terms.

● Public Domain Databases: Some platforms offer databases of works explicitly in the public domain or under open licenses. AI developers can tap into these resources without the need for additional licensing.

5. Micro-licensing Platforms: Emerging platforms use blockchain technology to facilitate micro-transactions and licenses. AI developers could use such platforms to automatically negotiate and pay for licenses when accessing copyrighted works.

6. Fair Use Licensing: In jurisdictions with a strong fair use doctrine, there might be a case for considering AI training as a form of transformative use. This approach would depend heavily on judicial interpretation and might not be a reliable long-term solution.

7. AI-Assisted Licensing: Interestingly, AI itself could be employed to facilitate the licensing process. AI algorithms could scan datasets, identify copyrighted material, and automatically send licensing requests to copyright owners or their representatives.

8. Custom Licenses for AI: Given the unique nature of AI training, there's potential for crafting custom licenses tailored to this purpose. These licenses could address issues like data retention, model sharing, and post-training use, providing clarity to both developers and copyright owners.

Conclusion: The evolving nature of AI presents both challenges and opportunities in the realm of copyright licensing. A balanced approach, which respects the rights of copyright owners while facilitating AI advancement, is essential. Whatever mechanisms are employed, they must be adaptable, transparent, and ensure that creators are fairly remunerated for their contributions.

10.1. Is direct voluntary licensing feasible in some or all creative sectors?

Direct voluntary licensing involves a copyright holder granting permission to another party to use their work under mutually agreed-upon terms. This method is straightforward and is founded on direct negotiations between the rights holder and the entity seeking to use the copyrighted content. Assessing its feasibility across various creative sectors requires a granular approach:

1. Music Industry:

● Feasibility: Moderate to High.
● Rationale: The music industry already employs direct licensing in various contexts, such as synchronization licenses for film and TV. Major artists or their representatives negotiate with entities desiring to use their work. However, when dealing with vast volumes of songs for AI training, direct licensing may become cumbersome unless streamlined through intermediaries.

2. Film and Television:

● Feasibility: Moderate.
● Rationale: Licensing in the film and TV sectors tends to be complex due to the number of stakeholders involved (actors, directors, composers, etc.). However, for specific projects or larger studios with established relationships, direct licensing can be effectively managed.

3. Publishing (Books, Articles, etc.):

● Feasibility: Low to Moderate.
● Rationale: The vast number of authors, especially in the academic domain, makes direct licensing challenging. Large publishers, however, might engage in direct negotiations for specific projects.

4. Visual Arts (Photography, Paintings, etc.):

● Feasibility: High for individual works, Low for large datasets.
● Rationale: For individual pieces, artists or their galleries can engage indirect licensing. But when AI developers need diverse and vast datasets, this method becomes less practical due to the sheer number of negotiations required.

5. Software and Gaming:

● Feasibility: Moderate to High.
● Rationale: Software licenses are often directly negotiated, especially for enterprise solutions. In the gaming world, licensing agreements for assets, characters, or game engines can also be directly brokered between parties.

6. Architecture and Design:

● Feasibility: Moderate.
● Rationale: Architectural designs, blueprints, or innovative designs can be directly licensed for specific projects or collaborations. However, for broad AI training on design principles or styles, direct negotiations might be less feasible due to the diversity of sources.

Conclusion: Direct voluntary licensing's feasibility varies across sectors, mainly depending on the number of stakeholders and the volume of content required. While effective for specific projects or high-value content, direct negotiations might not always be the most efficient approach for AI training that requires vast and diverse datasets. However, in sectors where relationships and collaborations are prevalent, this method holds significant promise.

10.2. Is a voluntary collective licensing scheme a feasible or desirable approach? Are there existing collective management organizations that are well-suited to provide those licenses, and are there legal or other impediments that would prevent those organizations from performing this role? Should Congress consider statutory or other changes, such as an antitrust exception, to facilitate negotiation of collective licenses?

Collective licensing, a system where rights holders band together to form an organization that negotiates rights on their behalf, offers a potential solution to the challenges posed by AI's extensive data requirements. Here's a closer look at its feasibility, existing structures, and potential impediments:

Feasibility and Desirability:

Efficiency: A collective licensing scheme can provide a streamlined solution for AI developers, enabling them to negotiate with a single entity rather than navigating agreements with countless individual rights holders.

Fair Compensation: Collective organizations can ensure that creators receive fair and transparent remuneration for the utilization of their works. This is particularly beneficial in sectors where individual artists or creators may not have significant bargaining leverage.

Flexibility: A collective approach can be agile, introducing tiered pricing or differentiated licenses based on the specifics of AI usage.

Existing Collective Management Organizations (CMOs):

Music Industry: Organizations such as BMI, ASCAP, and PRS manage music royalties. Their established infrastructures handle large volumes of copyrighted works and ensure the distribution of due royalties.

Publishing: Entities like the Copyright Clearance Center (CCC) are in place to oversee rights for various printed materials.

Visual Arts: Groups like VAGA and ARS cater to the rights of visual arts.

While these organizations function efficiently within their sectors, the expansive and diverse requirements of AI present unique challenges. It remains to be seen if existing CMOs can adapt to this new landscape or if the establishment of a specialized entity for AI's needs would be more effective.

Legal and Operational Impediments:

Diverse Needs: Given the multifaceted nature of AI applications, a single CMO might grapple with delivering all-encompassing solutions.

Global Reach: The international character of AI datasets introduces complexities in cross-border rights management. This raises concerns about the universal applicability of licenses.

Representation Concerns: Guaranteeing that all creators, regardless of their scale, are aptly represented within a vast CMO is a pressing challenge.

Congressional Involvement:

Considering the potential antitrust challenges linked with collective bargaining in a domain as influential as AI, congressional intervention may be essential. Potential areas of action include:

Antitrust Exemptions: Enabling collective negotiations without breaching antitrust regulations.

Regulatory Oversight: Instituting mechanisms to ensure that CMOs operate with transparency and fairness, safeguarding both AI developers and rights holders.

Standardized Licenses: Congress might advocate for or endorse standardized licensing frameworks to streamline the process and establish industry standards.

Conclusion:

Collective licensing emerges as a promising avenue to address some of AI's copyright challenges. However, translating this promise into practice is intricate and would likely demand a cooperative effort between industry stakeholders and legislative entities.

10.3. Should Congress consider establishing a compulsory licensing regime? If so, what should such a regime look like? What activities should the license cover, what works would be subject to the license, and would copyright owners have the ability to opt out? How should royalty rates and terms be set, allocated, reported and distributed?

A compulsory licensing regime for AI would mandate copyright holders to license their works to AI developers at a set rate. The thought behind this approach is to simplify the licensing process and guarantee AI developers access to the materials they need. However, implementing such a system would come with its own set of challenges and considerations.

Advantages:

- Uniform Access: Ensures that all AI developers, regardless of their size or resources, can access the data they need.
- Predictable Costs: Developers would know in advance the costs associated with using copyrighted material, aiding budgeting and cost projections.
- Efficiency: By passes the need for individual negotiations, speeding up development processes.

Designing a Compulsory Licensing Regime:

● Activities Covered: The license should cover activities integral to AI training, including data ingestion, preprocessing, model training, and validation. Depending on the AI's end use, some allowances for commercial outputs might also be considered.
● Scope of Works: Ideally, all copyrighted works should fall under this regime to provide the broadest utility. However, exceptions might be made for works of exceptional value or sensitivity.
● Opt-Out Mechanism: While the essence of a compulsory license is its universal applicability, provisions could be considered for rights holders to opt-out if they have strong objections, though this would complicate the system.

Setting and Managing Royalties:

● Rate Setting: Royalty rates could be based on the work's market value, its utility to AI developers, or a combination of both. Independent committees or industry consultations could help establish these rates.
● Allocation: Considering the sheer volume of data AI systems process, a per-use fee might be prohibitive. A tiered system based on the volume or type of data used could be more feasible.
● Reporting and Distribution: A centralized body, potentially an extension of an existing copyright organization, could manage the collection and distribution of fees. Regular audits and transparent reporting would ensure fairness and prevent misuse.
● Distribution: Royalties could be disbursed based on the frequency a particular work is used or its perceived value in the training set.

Potential Pitfalls:

● Setting Fair Rates: Determining a one-size-fits-all rate that's fair for all involved is a significant challenge.

● Administrative Overhead: The management of such a system would be complex, requiring a robust administrative mechanism.

● International Considerations: Given the global nature of AI development, there would be challenges related to international copyright laws and royalty distributions.

Conclusion:

While a compulsory licensing regime could significantly streamline AI's access to training data, its implementation would be fraught with challenges. Balancing the interests of rights holders with the needs of AI developers and ensuring a fair, transparent system would require careful planning, broad consultation, and likely, iterative refinement. The involvement of Congress in setting out the legislative framework and providing oversight would be crucial to its success.

10.4. Is an extended collective licensing scheme (51) a feasible or desirable approach?

Extended Collective Licensing (ECL) allows a collective management organization (CMO) to license works on behalf of all rights holders in a particular category, not just its members. In the realm of AI, adopting an ECL scheme might be a way to streamline the acquisition of training data, but it brings with it a set of advantages and challenges.

Advantages:

● Comprehensive Access: ECL provides AI developers with access to abroad range of copyrighted materials without negotiating individual licenses.
● Reduced Transaction Costs: Developers can secure rights from a single entity rather than myriad individual rights holders.
● Fair Remuneration: Rights holders are compensated even if they're not directly affiliated with the collective, ensuring a broader distribution of royalties.

Concerns and Challenges:

● Representation: An ECL scheme assumes that the collective accurately represents the interests of all rights holders in the category, which might not always be the case.
● Setting Rates: Like compulsory licensing, determining fair rates that satisfy both AI developers and rights holders could be contentious.
● Opt-Out Mechanisms: Rights holders should have a clear, easily accessible mechanism to opt out of the ECL if they do not wish their works to be included.
● Scope and Duration: The scope (which works are covered) and duration (how long the scheme is effective) must be clearly defined to prevent misuse and ensure rights holders' interests are protected.

Desirability and Feasibility:

● Sector Variation: ECL might be more suitable for certain sectors than others. For instance, in sectors where there's already a culture of collective management (like music or publishing), ECL might be more readily accepted and easier to implement.
● Existing Infrastructure: If a robust CMO is already in place, transitioning to an ECL scheme might be more straightforward. However, setting up a new organization for this purpose would entail significant logistical and administrative challenges.
● International Considerations: As with compulsory licensing, global AI development means that any ECL scheme would need to navigate international copyright laws and practices.

Conclusion:

ECL offers a promising mechanism for simplifying the licensing landscape for AI training data. It merges the comprehensiveness of compulsory licensing with the flexibility of voluntary collective licensing. However, its success hinges on transparent, fair representation of all rights holders and clear delineation of its scope and terms. Ensuring that an ECL system is both equitable and efficient will require collaboration between AI developers, rights holders, and potentially legislative bodies.

10.5. Should licensing regimes vary based on the type of work at issue?

The nature and nuances of different types of copyrighted works suggest that a one-size-fits-all approach to licensing for AI training might not be the most appropriate or efficient. Instead, tailoring licensing regimes based on the specific type of work in question could address unique challenges and opportunities presented by each category. Here's why and how this differentiation could be implemented:

Rationale for Differentiated Licensing:

● Varied Sensitivities: Different works carry varied cultural, moral, and economic sensitivities. For instance, using a book's text might not be seen in the same light as using personal photographs or sensitive indigenous art.
● Economic Disparities: The economic value and potential for monetization can vary drastically between, say, a popular song and a niche research article. Licensing regimes should account for these disparities to ensure fair compensation.
● Inherent Use Differences: An AI might use a music track differently from how it uses a piece of visual art or written content. The unique characteristics of each media type could dictate different licensing terms.

Potential Approaches:

● Segmented Licenses: Create distinct licensing categories for text, music, visual arts, films, and other media types. Each category would have its terms tailored to the typical uses and concerns associated with that medium.

● Variable Pricing: Adjust licensing fees based on the work's commercial potential, rarity, and cultural significance. For example, a historical painting might have a different fee structure than a stock photo.
● Duration-Based Terms: The length of time an AI system might require access to certain works could vary. Short-term access for specific projects might have different terms than long-term, continuous training.

Challenges:

● Complexity: Introducing multiple licensing regimes increases complexity for both AI developers seeking licenses and rights holders granting them. An overly complicated system might stifle innovation or lead to non-compliance.
● Defining Boundaries: Clearly demarcating categories and determining which works fall into which category can be a challenge, especially for works that straddle multiple mediums.
● Consistency: While differentiation is essential, maintaining some level of consistency across regimes will be crucial to ensure that the system is understandable and accessible.

Conclusion:

While a differentiated approach based on the work type seems logical and fair on the surface, its implementation requires a delicate balance. Striking the right compromise between specificity and simplicity, while ensuring that all stakeholders' concerns are addressed, is the key to a successful, tailored licensing regime. This approach would aim to recognize the distinct value and significance of each work type while facilitating the continued growth and innovation in the AI sector.

11. What legal, technical or practical issues might there be with respect to obtaining appropriate licenses for training? Who, if anyone, should be responsible for securing them (for example when the curator of a training dataset, the developer who trains an AI model, and the company employing that model in an AI system are different entities and may have different commercial or noncommercial roles)?

The act of obtaining appropriate licenses for AI training is a multifaceted endeavor, with challenges stemming from both legal and technical dimensions, as well as practical considerations.

Legal Challenges:

● Ambiguity in Current Laws: Existing copyright laws might not explicitly cover AI-related activities, making it uncertain what requires licensing.
● Jurisdictional Differences: Different countries may have varied approaches to copyright and AI, complicating licensing for global operations.
● Multiplicity of Stakeholders: A piece of content might have multiple rights holders, making licensing more complex.
● Determining Liability: It's unclear who should be held responsible if unlicensed content finds its way into training datasets.

Technical Challenges:

● Scale: AI often requires vast datasets, making manual licensing impractical.
● Identifying Copyrighted Material: It might be challenging to determine if a piece of data in a vast dataset is copyrighted, especially if metadata is missing or unclear.
● Dynamically Updated Datasets: Datasets that update in real-time pose a challenge as new data might introduce unlicensed content.
● Traceability: Once an AI model is trained, it's difficult to pinpoint which data it was trained on, complicating compliance verification.

Practical Challenges:

● Economic Feasibility: Licensing vast amounts of data individually might be economically unviable for many AI developers.
● Time Constraints: The pace of AI development often requires quick access to data, and waiting for licensing approvals can hinder progress.
● Public Domain Assumptions: Many might assume that older works or widely available datasets are in the public domain, leading to unintentional copyright breaches.

Responsibilities:

● Dataset Curators: Those who assemble and distribute datasets bear primary responsibility for ensuring the data they provide is either licensed, license-exempt, or in the public domain.
● AI Developers: Developers using datasets should verify the legitimacy of the dataset source and confirm appropriate licensing, especially if making commercial use of the trained model.
● Employing Company: Companies that deploy AI models in their systems should have a due diligence process to ensure that the model's training complies with copyright laws. They might also be responsible for obtaining licenses for further distribution or commercial applications.
● Third-Party Verification: There could be a role for independent entities to certify datasets or AI models as compliant with licensing norms.

Conclusion:

Given the intricate landscape of AI training licensing, collaboration among stakeholders is imperative. Clear guidelines, possibly with legal revisions, can pave the way for more straightforward licensing processes. It's also crucial to assign clear responsibilities across the AI training pipeline to ensure that all involved parties understand their roles in maintaining copyright compliance.

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

Understanding the precise contribution of a particular work to the output of a generative AI system is a complex issue due to the intricacies of AI's operational mechanisms. Here's a breakdown of the challenges and feasibility:

Inherent Complexity of AI Models:

● Black Box Nature: Many AI models, especially deep neural networks, operate as "black boxes," meaning it's challenging to ascertain how they make decisions or produce outputs.
● Interwoven Knowledge: AI models trained on vast datasets digest information collectively. A single input contributes to the model in conjunction with all other inputs, making it hard to isolate its singular impact.

Technical Challenges:

● Distributed Representations: In neural networks, knowledge isn't stored in a singular neuron or layer. Instead, it's distributed across the network. This distribution means that the influence of any one piece of data is diffused and intertwined with all other data points.
● Non-linearity: Many AI models introduce non-linear transformations, causing intricate relationships between inputs and outputs. Tracing back the influence of a single input becomes increasingly difficult due to this non-linearity.

Practical Challenges:

● Vast Datasets: Given the enormous size of training datasets, even if it were technically possible to measure the influence of a single data point, doing so for every piece of data would be practically infeasible.
● Transient Training Influence: As models undergo further training or fine-tuning on new data, the influence of earlier data points might diminish or evolve.

Feasibility:

● Contribution Metrics: Some methodologies, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), attempt to provide insights into how different features influence predictions in specific models. However, these are more geared towards understanding features rather than the influence of individual training examples.
● Counter factual Analysis: This involves assessing model outputs with and without specific pieces of training data. While theoretically possible, this would be computationally intensive and might still not offer precise measures of contribution.

Conclusion:

Given the current state of AI and machine learning, it is extremely challenging, if not impossible, to accurately quantify the degree to which a specific work contributes to a particular AI-generated output. While some methods provide insights into model decisions, a granular understanding of individual data point contributions remains elusive.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

The introduction of a licensing requirement for copyrighted materials used in training generative AI systems could have significant economic repercussions. Here's a detailed breakdown of potential impacts:

1. Cost Implications:

● Immediate Costs: AI developers and firms would need to secure licenses for all copyrighted material they use, leading to higher costs upfront.
● Administrative Burden: Managing, tracking, and renewing licenses, especially if numerous sources are involved, would entail additional administrative costs.

2. Innovation and R&D:

● Potential Stifling: The added costs and complexities could deter startups and smaller firms from venturing into generative AI, possibly slowing down innovation in the field.
● Shift in Focus: Companies might shift their R&D focus to areas where they can easily obtain or do without copyrighted content, altering the direction of AI advancements.

3. Open-Source and Collaborative Research:

● Barrier to Entry: Licensing requirements could serve as barriers to entry for open-source projects or academic research, where budgets are limited, and the spirit is collaborative.

4. Quality and Diversity of Training Data:

● Compromised Quality: Developers might opt for easily accessible or affordable data over high-quality copyrighted data, potentially affecting the performance and capabilities of the resultant AI models.
● Less Diverse Models: Limited access to diverse copyrighted content might result in models that are less versatile or that perpetuate existing biases.

5. Market Dynamics:

● Consolidation: Larger companies with more resources could better navigate the licensing landscape, potentially leading to market consolidation at the expense of smaller players.
● Emergence of Data Brokers: A new market for third-party data brokers or aggregators might emerge, who curate and offer licensed datasets specifically for AI training.

6. Global Competitiveness:

● Shift in AI Leadership: Countries with more flexible or AI-friendly licensing frameworks might attract more AI research and business, possibly shifting the global leadership in AI innovation.
● Harmonization Challenges: Cross-border AI projects might face challenges in harmonizing licensing requirements of different countries.

7. Commercial Applications:

● Increased Product Costs: The costs associated with licensing might be passed on to consumers in the form of higher prices for AI-driven products and services.
● Limited Offerings: Some AI-driven solutions that rely heavily on copyrighted data might become economically unviable and be withdrawn from the market.

Conclusion:

While licensing requirements are intended to protect the rights of copyright holders, they could introduce significant economic challenges for the AI sector. Balancing the rights of content creators with the potential of AI will be crucial to ensure that both can thrive in this evolving landscape.

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.

1. Evolution of AI Technology:

● Rapid Advancements: AI and machine learning are rapidly evolving fields. As technology becomes more advanced, the ways in which copyrighted content is used and processed by AI models may change, potentially complicating copyright considerations.

2. Inherent Nature of AI Training:

● Transformational Usage: AI models typically use copyrighted materials in a transformational manner, converting raw data into patterns and abstract representations. This non-literal use could be a significant factor in assessing copyright infringements.

3. Public Benefits:

● AI's Broader Impact: The wider societal and economic benefits brought by AI — such as medical advancements, better disaster response, or improved accessibility services — may be weighed against the strict enforcement of copyright liabilities.

4. Ambiguities in Identifying Infringement:

● Opaque Processes: Neural networks, especially deeplearning models, are often termed as "black boxes" due to their opacity. Determining whether a model has infringed copyright based on its outputs can be ambiguous given the complexity of these processes.

5. Historical Analogies:

● Prior Technologies: Historical precedents, like the use of copyrighted material in search engines or digital libraries, could provide insights into how courts and policymakers might approach copyright concerns with AI training.

6. Ethical Considerations:

● Bias and Representation: Ensuring that AI models are unbiased and representative may require diverse training data. Restrictive copyright practices might impede this, raising ethical issues around the fair and equitable use of AI technologies.

7. Precedents in Other Jurisdictions:

● Global Approaches: How other countries tackle copyright issues in AI training can offer valuable insights, especially given the global nature of AI research and development.

8. Ephemeral Copies and Caching:

● Temporary Reproductions: AI training often involves temporary reproductions of data. The treatment of such ephemeral copies, in terms of copyright, can be a relevant factor, similar to how caching is treated in web technologies.

9. Non-economic Uses:

● Educational and Research Purposes: AI models used for academic, research, or other non-commercial purposes might be considered differently in terms of copyright liabilities.

10. Digital Rights Management (DRM):

● Technical Protections: The interaction between AI training and DRM systems, which are designed to prevent unauthorized use of copyrighted materials, could introduce another layer of complexity.

Conclusion:

The copyright landscape for AI training is multifaceted, with technological, societal, ethical, and historical factors all playing roles. As AI continues to become more ingrained in our daily lives, it's essential to consider these various factors holistically to strike the right balance between innovation and protection of intellectual property.

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation?

Record-keeping and Disclosure Requirements for AI Model Developers and Dataset Creators

Pros:

1. Accountability and Transparency:

● Requiring AI developers and dataset curators to maintain records ensures accountability. It makes it possible to ascertain whether copyrighted works have been used without authorization.

2. Facilitates Licensing:

● With clear records, it becomes easier for copyright owners to negotiate licensing agreements, knowing that their works are being used.

3. Protection against Infringement:

● If AI developers and dataset creators maintain transparent records, they can avoid potential litigation by proving they haven't used any unauthorized copyrighted materials.

4. Encourages Ethical Practices:

● Mandatory disclosure may promote ethical data collection practices, reducing the chances of using datasets that have bias or other unethical content.

5. Enables Audits:

● Clear record-keeping will allow for third-party audits, ensuring that AI models are being trained on ethically sourced, non-infringing data.

Cons:

1. Impracticality:

● Given the vast amount of data required to train some AI models, maintaining exhaustive records might be impractical. Some datasets contain millions or even billions of data points.

2. Privacy Concerns:

● Mandatory record-keeping might infringe on privacy rights, especially if the datasets contain personal information.

3. Increased Costs:

● Storing, managing, and potentially disclosing records can increase the operational costs for AI developers and dataset creators.

4. Stifling Innovation:

● Overly stringent requirements might deter small-scale developers or startups from venturing into AI development due to the administrative burden.

5. Proprietary Information:

● Requiring disclosure of training materials might force companies to reveal proprietary datasets or methods, which could compromise competitive advantages.

6. Ambiguity in Disclosure:

● Deciding what level of detail should be disclosed can be ambiguous. It could be challenging to strike a balance between comprehensive disclosure and overwhelming granularity.

Conclusion:

While the idea of record-keeping and disclosure is rooted in ensuring transparency and protecting copyrights, practical implementation would require careful consideration. Striking a balance between copyright protection and fostering innovation is crucial. Establishing a clear and standardized framework, perhaps with varying requirements based on the nature and scale of AI development, could provide a solution that serves both creators and AI developers.

15.1. What level of specificity should be required?

Level of Specificity in Record-Keeping and Disclosure for AI Model Developers and Dataset Creators

When considering the level of specificity required for record-keeping and disclosure, it's important to balance the granularity of details with the practicality of the process. The primary objectives should be transparency, accountability, and ensuring that copyright laws are not violated. Here's an examination of potential levels of specificity:

1. Basic Level:

● Details: Name of the dataset, source (if publicly available), and broad categorization of data (e.g., images, texts, audio).
● Pros: Simplified record-keeping, less administrative overhead.

● Cons: Might not be sufficient for copyright holders to determine if their work has been used.

2. Intermediate Level:

● Details: Includes all from the basic level, plus data acquisition method (scraped, purchased, user-generated, etc.), number of data points, and any licensing or permission details.
● Pros: Offers a better understanding of data origins and potential legal compliance.
● Cons: More administrative work than the basic level but might still lack detail for specific copyright checks.

3. Detailed Level:

● Details: Incorporates all from the intermediate level, with a thorough breakdown of datasets into sub-categories, sample data points, and any transformation or pre-processing steps taken on the data.
● Pros: Comprehensive insight into the dataset, facilitating copyright verification.
● Cons: High administrative overhead, potential risks related to privacy, and concerns over revealing proprietary information.

4. Exhaustive Level:

● Details: Includes all previous levels, along with exhaustive lists of every single data point, metadata, data sources, timestamps, and more.
● Pros: Maximum transparency.
● Cons: Impractical for vast datasets, high storage and management costs, heightened privacy and proprietary risks.

Recommendation:

An intermediate to detailed level of specificity strikes a reasonable balance for most scenarios. AI developers could maintain detailed internal records but disclose at an intermediate level, providing more granularity upon specific requests or challenges by copyright holders. This approach ensures accountability while protecting privacy and proprietary information. Additionally, a tiered system could be implemented, adjusting the specificity based on the type of data, potential for copyright infringement, and the scale or purpose of the AI model.

15.2. To whom should disclosures be made?

The audience for disclosures will vary based on the objective of the disclosure, the nature of the AI system, and the regulatory environment. Here's a breakdown of potential recipients and the reasons for each:

1. Copyright Authorities or Regulatory Bodies:

● Purpose: Oversight and enforcement of copyright laws.
● Details: Regulators would be the primary point of contact for overseeing the compliance of AI models with copyright regulations. They can intervene if there's a violation or a complaint.

2. Copyright Owners or Stakeholders:

● Purpose: Verification and potential remuneration.
● Details: If there's a suspicion or evidence that copyrighted material has been used without authorization, copyright owners should be granted access to relevant records to verify claims.

3. Public Domain (Open Disclosure):

● Purpose: Transparency and public accountability.
● Details: Particularly for AI models developed or utilized by public entities or those serving public functions, an open disclosure could ensure transparency and trust.

4. Third-Party Auditors:

● Purpose: Neutral assessment and verification.
● Details: Independent organizations could verify compliance without bias, ensuring the protection of proprietary information while confirming adherence to regulations.

5. Licensing Bodies or Collective Management Organizations (CMOs):

● Purpose: Management of permissions and royalties.
● Details: If AI developers are sourcing copyrighted material through licensing bodies or CMOs, these organizations may require disclosure to verify adherence to licensing agreements and to distribute royalties appropriately.

6. Courts or Legal Entities:

● Purpose: In the event of disputes.
● Details: In cases of legal disputes over copyright infringement, courts or relevant legal entities might require detailed disclosure to resolve the matter.

7. Partners or Collaborators:

● Purpose: Collaboration and joint projects.
● Details: In joint AI projects, co-developers, partners, or collaborators may need access to training data records to align their efforts and ensure mutual compliance.

Recommendation:

The nature of the AI system, its application, and the potential risks associated with its training data will determine to whom disclosures should be made. For systems with a broad public impact or those in regulated sectors, a higher degree of transparency might be warranted. However, care should always be taken to ensure that proprietary, personal, or sensitive information is adequately protected in any disclosure process.

15.3.What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?

When developers incorporate third-party AI models into their systems, they inevitably inherit a degree of responsibility related to the integrity, legality, and performance of those models. Here are some proposed obligations:

1. Due Diligence:

● Developers should conduct a thorough vetting of third-party models to ensure they adhere to legal and ethical standards. This includes understanding the origin of the training data and ensuring no copyright infringements or other legal issues are present.

2. Transparency and Disclosure:

● Developers should provide clear information about the third-party models they use, their origins, and any relevant certifications or attestations they possess.

3. Update and Maintenance:

● If a third-party model is found to have issues after its integration, developers should take the initiative to rectify the problem, whether by updating, replacing, or discontinuing use.

4. Liability and Accountability:

● While third-party developers bear primary responsibility for their models, those incorporating these models into broader systems should also assume some level of liability, especially if they neglect due diligence or knowingly use flawed models.

5. Data Privacy and Security:

● Developers must ensure that third-party models adhere to data privacy regulations and standards. They should also confirm that any data used to fine-tune or further train these models are handled securely and ethically.

6. Communication with Third-Party Developers:

● Maintain an open line of communication with third-party model developers. This ensures prompt updates on potential issues, improvements, or changes to the model that could affect its performance or legality.

7. Right to Audit:

● In licensing or agreement terms, developers could negotiate a 'right to audit' clause, granting them permission to review the processes and data behind third-party models, ensuring ongoing compliance and quality.

8. Public Representation:

● Developers should avoid misrepresenting third-party models as their own and should provide appropriate attribution or credits where necessary.

9. Contingency Planning:

● Developers should have a strategy in place for scenarios where a third-party model is found to be problematic or becomes unavailable. This could involve having backup models, modular system designs, or alternative data sources.

10. Feedback Mechanism:

● Developers should have a mechanism to receive feedback about potential issues or concerns related to third-party models and take appropriate action in response.

Recommendation:

While third-party AI models can offer efficiency and expertise, developers incorporating these models should be proactive in ensuring their quality, legality, and ethical standing. Balancing trust with verification will be key in maintaining high standards while benefiting from external innovations.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

Impact of a Recordkeeping System on Various Stakeholders

Implementing a recordkeeping system for developers of AI models or systems presents both costs and benefits. The impact of such a system varies for different stakeholders:

1. Developers of AI Models or Systems:

Costs:

- Infrastructure: Developers might need to invest in servers, databases, and other hardware to store records.
- Maintenance: Ongoing costs for maintaining, backing up, and securing the database.
- Labor: Hiring or reallocating personnel for data entry, validation, and management.
- Training: Ensuring that personnel are trained to correctly use and maintain the system.
- Legal & Compliance: Navigating regulations, addressing disputes, and potential liability from mismanagement.

Benefits:

● Transparency and Trust: Demonstrates commitment to ethical practices, potentially winning trust from clients and partners.
● Dispute Resolution: Having detailed records can expedite the resolution of copyright or data use disputes.

2. Creators:

Costs:

● Time: It might take time for creators to verify or check the use of their works in the system.
● Privacy Concerns: Potential exposure of certain work details or other related metadata.

Benefits:

● Control: Allows creators to monitor the usage of their content and take action if necessary.
● Monetization: If licensing fees are applicable, a transparent record can facilitate fair compensation.

3. Consumers:

Costs:

● Potential Cost Increase: The overhead costs of the recordkeeping system might be passed onto consumers in the form of higher prices for AI products or services.

Benefits:

● Transparency: Consumers can understand the provenance of AI models and systems, allowing for informed choices.
● Trust: Knowing that developers adhere to ethical and transparent practices can enhance trust in AI products.

4. Other Relevant Parties (e.g., Regulators, Researchers):

Costs:

● Oversight and Enforcement: Regulators may need to invest in mechanisms to oversee and enforce compliance.

Benefits:

● Insight and Understanding: Access to records can provide valuable insights for research and policy formulation.
● Standardization: Helps in setting industry standards for AI training and model development.

Overall Consideration:

While there are evident costs associated with the implementation of a recordkeeping system, the long-term benefits, particularly in fostering transparency, trust, and standardization, can outweigh these costs. However, it's essential to consider scalability, especially for startups or small developers, to ensure that they aren't disproportionately burdened.

16. What obligations, if any, should there be to notify copyright owners that their works have been used to train an AI model?

Obligations to Notify Copyright Owners about AI Model Training Usage

When considering the obligations related to notifying copyright owners about their works being used for training AI models, several factors come into play. Balancing the interests of copyright owners and AI developers is paramount. Here's a proposed framework:

● Nature of Use:
○ *Commercial Use*: If the AI model is intended for commercial purposes, there should be an obligation to notify the copyright owner. This would ensure that creators are made aware when their works potentially contribute to profit-driven ventures.
○ *Research and Non-Commercial Use*: For academic research or other non-commercial endeavors, the obligation might be more relaxed, given the non-profit nature of the project.
● Volume and Magnitude of Use:

◯ If a work plays a significant role in the training data or has a profound impact on the resulting model, notifying the copyright owner becomes crucial. On the other hand, if the copyrighted material is just one among millions of data points and does not profoundly influence the model's behavior, the obligation might be more lenient.

● Mechanism for Notification:

◯ A centralized repository or database could be established where developers can list copyrighted materials used in AI training. Copyright owners can then periodically check this repository.

◯ Alternatively, direct notifications (via email or other digital means)could be sent to copyright owners when their works are incorporated into new AI models.

● Opt-Out Provisions:

◯ Alongside the notification, there should be a mechanism that allows copyright owners to request the removal of their work from the training set, especially if they disagree with the usage or the context in which it's used.

● Exceptions and Special Cases:

◯ Some materials, like open-source content, public domain works, or those under specific licenses like Creative Commons, may have different notification requirements based on their licensing terms.

◯ Considerations should also be made for legacy works where the copyright owner might be unreachable or unknown.

● Practical Implications:

◯ The feasibility of notification will depend on the size and nature of the AI developer's operations. Large tech companies might have the infrastructure to notify copyright owners systematically, but this could be burdensome for smaller entities or individual researchers. Any obligation imposed should account for these disparities.

● Legal and Jurisdictional Issues:

◯ Copyright laws vary across countries. A work might be copyrighted in one jurisdiction and not in another. This complexity should be considered when determining notification obligations.

In essence, while there's a genuine interest in ensuring that creators are recognized and can control the usage of their works, practical, scalable, and efficient methods must be devised to notify copyright owners without stifling innovation in the AI domain.

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?

U.S. Laws Potentially Impacting AI Model Training Disclosure Outside of Copyright Law

Outside of copyright law, there are several U.S. legal frameworks that, either directly or indirectly, could impose requirements on developers of AI models or systems concerning the retention or disclosure of training materials. Some of the pertinent areas of law include:

● Trade Secrets: Under the federal Defend Trade Secrets Act (DTSA) and corresponding state laws, companies may wish to keep their training data confidential as a trade secret. If litigation arises about misappropriation, the details of the training data and methods may need to be disclosed, albeit under protective court orders.
● Consumer Protection Laws: The Federal Trade Commission Act prohibits deceptive and unfair practices. If an AI system impacts consumers, and its behavior can be traced back to its training data, details about that training may be relevant in regulatory or legal actions concerning deceptive practices.
● Data Privacy Laws: Regulations such as the California Consumer Privacy Act (CCPA) and its successor, the California Privacy Rights Act (CPRA), have stipulations regarding the collection, usage, and disclosure of personal information. If AI training data includes personal information of California residents, these laws could impose obligations to disclose the sources of that information and its intended use.

● Healthcare: In the context of healthcare, the Health Insurance Portability and Accountability Act (HIPAA) sets standards for the protection of patient health information. If AI is trained using patient data, disclosures about the source and nature of that data might be required, particularly if there's a breach.
● Financial Services: For AI used in the financial sector, various regulations might come into play. For instance, under the Dodd-Frank Wall Street Reform and Consumer Protection Act, there are stipulations for transparency and accountability in decision-making processes, which might indirectly touch on AI training data.
● Federal Acquisition Regulations (FAR): For companies developing AI systems under contracts with the federal government, FAR might require certain disclosures about the technology's development, potentially including training data specifics.
● National Security and Export Control: U.S. export control laws can impose restrictions on the export of specific AI technologies and might require declarations or disclosures about the technology's nature, which could indirectly relate to training data.
● Algorithmic Accountability: While not yet codified into law, there have been discussions and proposed bills at both federal and state levels about algorithmic transparency and accountability. Such laws, if passed, would likely require entities to disclose how AI systems operate, which might necessitate details about training data.

While many of these laws don't directly require the retention or disclosure of AI training materials, they create environments where such disclosure might be necessary or advantageous. As AI continues to evolve and integrate into various sectors, it's conceivable that more specific regulations will emerge to address these and related issues.

18. Under copyright law, are there circumstances when a human using a generative AI system should be considered the “author” of material produced by the system? If so, what factors are relevant to that determination? For example, is selecting what material an AI model is trained on and/or providing an iterative series of text commands or prompts sufficient to claim authorship of the resulting output?

The question of whether a human can be considered the "author" of material produced by a generative AI system is a complex one and touches on fundamental tenets of copyright law. Here are some considerations:

- Human Creativity as the Bedrock of Copyright: At its core, copyright law is designed to protect original expressions of human creativity. Traditionally, for a work to be copyrighted, it needs to be original and the product of human authorship.
- Current Legal Framework: The U.S. Copyright Office's Compendium of U.S. Copyright Office Practices (Third Edition) states that it will register "an original work of authorship, provided that the work was created by a human being."
- Human Involvement and Iteration: If a human provides significant input, direction, or decision-making in the process, it's arguable that the final product is a collaborative effort between human and machine. For instance:
  - Training Data Selection: If a person curates the training data for the AI, they are making decisions that shape the AI's knowledge base and potential outputs. However, it's uncertain if this alone suffices for copyright claims since the output isn't directly authored by the human.
  - Prompts and Commands: Direct human guidance during the AI's creative process, such as giving specific commands or iterative feedback, could lend more weight to the argument that the human is a co-author of the output.
- Threshold of Originality: Even if a human can claim some level of authorship, the AI-generated work still must meet the threshold of originality. If the AI simply rearranges or slightly modifies existing copyrighted content, that output might not be considered original.
- Comparison with Other Collaborative Tools: It can be argued that using AI as a tool is analogous to using other instruments, software, or techniques in the creative process. For instance, a photographer uses a camera, an artist might use digital software, and a writer might use word prediction tools. Each of these tools aids the human, but doesn't negate their claim to authorship. The line, however, gets blurred when the tool (in this case, AI) becomes more autonomous in its creative contributions.

● Legal Precedents and Evolution: As AI becomes increasingly integrated into creative processes, courts and legislatures will likely grapple with these issues moredirectly.Newcasesorlegislationcouldprovideclearerguidanceon when and how human involvement in AI-generated works constitutes authorship.

19. Are any revisions to the Copyright Act necessary to clarify the human authorship requirement or to provide additional standards to determine when content including AI-generated material is subject to copyright protection?

The integration of AI into the creative landscape is undoubtedly transforming the way content is produced, which in turn raises complex questions about copyrightability. Given the advancements and increasing utilization of AI, there's a compelling argument for revisiting the Copyright Act to address these emerging challenges. Here are some potential areas of consideration:

● Explicit Definition of 'Authorship': The Copyright Act could be amended to include a clearer definition of what constitutes "human authorship" in the context of AI-assisted or AI-generated works. This could provide better guidance for determining when a work involves sufficient human intervention to merit copyright protection.
● Acknowledgment of AI-Assisted Works: The Act might introduce categories or classifications that recognize AI-assisted works, providing a framework to understand how these works differ from traditional creations and how they should be treated legally.
● Threshold for Originality: Given that AI can produce vast amounts of content rapidly, the Act could provide clearer guidance on what meets the "threshold of originality" in AI-generated contexts. This would help avoid flooding the copyright system with works that lack significant originality.
● Ownership of AI-Generated Works: If an AI-generated work is deemed copyrightable, the Act could clarify who holds the rights to it. Is it the developer of the AI software, the operator who provided inputs/prompts, or perhaps the entity that owns the AI system?
● Duration of Copyright for AI Works: Given the potential for AI to generate content at an unprecedented scale, the Act might consider whether

AI-generated or AI-assisted works should have a different copyright term than traditional works.
● Moral Rights Consideration: For works where AI plays a significant role but there's discernible human involvement, there might be questions about moral rights. This would involve understanding whether AI-generated alterations to human works might harm the reputation or integrity of the original human creator.
● Fair Use in Training AI: The Act could provide clarity on how the doctrine of fair use applies to the utilization of copyrighted materials in training AI models. Given the vast amounts of data AI models can process, traditional understandings of "amount and substantiality" could be revisited.
● Mechanisms for Licensing and Royalties: The Act could explore new models for licensing and royalties that accommodate the unique challenges posed by AI, such as collective licensing regimes or compulsory licenses tailored for AI training data.
● Transparency and Accountability: Regulations around transparency in AI training, including potential requirements for disclosure about training data sources, could be beneficial for copyright holders.
● International Consistency: Given the global nature of digital content and software development, considerations might be given to how U.S. copyright regulations regarding AI align or differ from international norms and treaties.

In conclusion, the intersection of AI and copyright is a dynamic and evolving area. It might be beneficial for legislators, legal experts, AI researchers, and stakeholders from the creative industries to collaboratively consider potential revisions to the Copyright Act, ensuring it remains relevant and effective in the age of AI.

20. Is legal protection for AI-generated material desirable as a policy matter? Is legal protection for AI-generated material necessary to encourage development of generative AI technologies and systems? Does existing copyright protection for computer code that operates a generative AI system provide sufficient incentives?

Desirability of Legal Protection for AI-Generated Material:

● Incentive for Development: Legal protection may stimulate investment in AI development by ensuring that the output of these systems can be protected and monetized.
● Clarification and Certainty: Legal clarity can prevent potential disputes and provide guidelines for industries seeking to leverage AI for content creation.
● Recognition of Effort: While AI generates the content, significant human effort is invested in designing, training, and fine-tuning these models. Protecting AI outputs could acknowledge these contributions.
● Potential Misuse and Over protection: On the flipside, offering extensive copyright protection for AI-generated content might lead to strategic copyrighting, where entities produce massive amounts of content merely to claim rights over broad domains.

Necessity for Encouraging Development:

● Sufficiency of Current Incentives: AI development is currently driven by a myriad of factors, including the potential for automation, efficiency gains, innovation in products/services, and competitive advantages. The promise of copyright protection for AI outputs might not be a primary driver.
● Potential Barriers: Imposing strict copyright regulations on AI-generated content could also impede the open-source ethos prevalent in the AI community, potentially slowing innovation.
● Diverse Applications: The importance of copyright might vary depending on the application. For artistic AI creations like music or art, copyright could be more pivotal than for other applications, such as data analysis.

Protection for Computer Code Operating AI:

● Distinguishing Code from Output: While the computer code that drives AI can be copyrighted, this protection doesn't extend to the myriad potential outputs the AI can generate. They are fundamentally different domains of creation.
● Incentive Alignment: Copyrighting AI code protects the intellectual property of the software developers, but it might not address the interests of those who employ AI for content generation.
● Limited Scope of Code Protection: Protecting the AI model's code might not provide comprehensive incentives for the broader ecosystem involved in AI

content generation, including those curating training data, refining model outputs, or integrating AI content into larger works.

In summary, while there are arguments in favor of extending legal protections to AI-generated content, such decisions should be balanced against potential impediments to innovation and the broader public interest. The nuances of AI's role in creation should be considered, and any legal framework should be flexible enough to adapt to rapidly advancing technology.

20.1. If you believe protection is desirable, should it be a form of copyright or a separate *sui generis* right? If the latter, in what respects should protection for AI-generated material differ from copyright?

Advantages of Using Copyright Protection:

● Established Framework: Copyright is a well-understood and established system of protection with defined rights, limitations, and an existing enforcement mechanism.
● International Recognition: Copyright treaties provide international standards for protection, making it easier to enforce rights globally.
● Integration with Existing Systems: Many platforms and industries are already designed to respect and handle copyright claims.

Disadvantages of Using Copyright Protection:

● Misfit with Traditional Principles: Copyright was designed with human creators in mind. AI outputs might not fit neatly within the established notions of creativity and originality.
● Duration Issues: The standard duration of copyright (lifeoftheauthorplus70 years in many jurisdictions) is irrelevant for AI-generated outputs.
● Unclear Ownership: Determining the "author" of AI-generated content could be complex, especially when multiple entities are involved in training and refining the AI.

Advantages of a Sui Generis Right:

● Tailored Protection: A separate right can be designed specifically for AI outputs, addressing unique challenges and considerations.

● Defined Duration: A sui generis system can have a specific duration tailored to the lifecycle of AI-generated content.
● Addressing Ownership and Licensing: Such a system could provide clear rules on ownership, licensing, and transfer of rights, taking into account the various entities involved in AI content generation.
● Balancing Public and Private Interests: A bespoke system could strike a balance between incentivizing AI development and ensuring public access to AI-generated content.

Disadvantages of a Sui Generis Right:

● Complexity: Introducing a new legal right could add complexity to IP law and create potential overlaps or conflicts with existing rights.
● International Challenges: A sui generis system might not be immediately recognized internationally, complicating cross-border enforcement.
● Potential Overprotection: If not carefully designed, a sui generis system could overprotect AI outputs, stifling derivative works and innovation.

In What Respects Should Protection Differ from Copyright?

● Duration: AI-generated content could have a fixed protection period, irrespective of "author" lifespan, perhaps shorter than traditional copyright to reflect the rapid evolution of technology.
● Threshold for Protection: Given the volume of content AI can generate, the threshold for protection might need to be higher, focusing on truly unique or value-added outputs.
● Rights Conferred: A sui generis system might limit certain exclusive rights or introduce mandatory licensing provisions to balance access and incentive.
● Moral Rights: Traditional copyright often recognizes moral rights, like the right to attribution. These might be less relevant or differently conceptualized for AI outputs.

In conclusion, while copyright offers a familiar framework, the unique nature of AI-generated content might warrant a specialized system. A sui generis right, if carefully crafted, can provide tailored protection that respects the interests of developers, users, and the public.

21. Does the Copyright Clause in the U.S. Constitution permit copyright protection for AI-generated material? Would such protection “promote the progress of science and useful arts”? If so, how?

Copyright Protection for AI-Generated Material under the U.S. Constitution's Copyright Clause

The Copyright Clause in the U.S. Constitution, Article I, Section 8, Clause 8, empowers Congress "To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries."

Does the Copyright Clause Permit Copyright Protection for AI-Generated Material?

● Authorship: The primary hurdle arises from the term "authors." Traditionally, an author is viewed as a human creator. If AI-generated material is to be copyrighted, a determination must be made about who, if anyone, constitutes the "author" — the human programmer, the user who commands the AI, or some other human actor. Without a clear human author, traditional copyright protection might be constitutionally problematic.
● Promotion of Progress: The Copyright Clause's primary objective is to promote progress. If it can be argued that granting copyright to AI-generated works incentivizes innovation and advances in AI technology, then such a provision might align with the spirit of the clause. However, the counter-argument is that such protection could stifle progress by excessively limiting the use and distribution of AI-generated outputs.

Would Such Protection “Promote the Progress of Science and Useful Arts”?

● Incentivizing Development: Providing copyright protection could incentivize companies and individuals to invest in the development of advanced AI systems, believing they can protect and monetize the AI's outputs.
● Boosting the AI Industry: With clear protective measures in place, businesses might feel more secure in integrating AI into various sectors, from entertainment to research, which could stimulate growth in AI applications.
● Balancing Interests: On the flipside, too much protection might hinder progress. If AI-generated content enjoys extensive protection, it might deter

derivative works, adaptations, or other creative endeavors that could otherwise emerge from unrestricted access to AI outputs.
● Encouraging Quality Over Quantity: There's a potential for AI to generate vast amounts of content quickly. Copyright protection could encourage the development of AI systems that generate high-quality, unique content rather than merely flooding the market with generic outputs.

Conclusion:

While the constitutional permissibility of extending copyright to AI-generated works remains a debated topic, the central question revolves around the balance between encouraging technological advancement and ensuring that the fruits of such advancements remain accessible and beneficial to society. Ultimately, the answer might lie in a middle ground, such as a modified protection system or a sui generis right that acknowledges the unique characteristics of AI outputs.

22. Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right? If so, in what circumstances?

AI-generated outputs can indeed implicate the exclusive rights of preexisting copyrighted works, especially in relation to the rights of reproduction and the creation of derivative works. Here are some scenarios in which this might happen:

Right of Reproduction

- Direct Reproduction: If an AI system is trained on copyrighted materials and its output directly reproduces substantial parts of those copyrighted materials, it would infringe upon the right of reproduction.
- Indirect Reproduction: In some cases, AI could produce outputs that, while not identical, are substantially similar to copyrighted works, potentially raising issues related to the right of reproduction.

Derivative Work Right

- Transformation: An AI system might generate content that incorporates elements of copyrighted works in a new context. If recognizable, these could be seen as unauthorized derivative works.

- Adaptation: Some AI systems can generate different types of media (text, audio, images). If an AI converts a copyrighted text into a piece of music that maintains the original work's expression, this could be seen as creating an unauthorized derivative work.

Circumstances to Consider

● Nature of the AI Model: The more an AI system is designed to generate outputs that closely resemble its training data, the higher the risk of infringement.
● Commercial vs. Non-commercial Use: Commercial use of AI-generated outputs that implicate copyrighted materials could result in higher penalties and is less likely to be considered fair use.
● Extent of Similarity: Legal scrutiny often involves an assessment of how similar the AI-generated work is to the copyrighted material. The more substantial the similarity, the higher the risk of infringement.
● Quantity of Infringing Material: If only a very small portion of the output can be traced back to copyrighted content, it may not constitute infringement, although this is highly contextual.
● Fair Use Doctrine: Depending on various factors such as the purpose of use, nature of the copyrighted work, amount used in relation to the work as a whole, and the effect on the market value, use of copyrighted material may qualify as fair use, but this is a complex legal determination.

In conclusion, AI-generated outputs can potentially implicate the exclusive rights of preexisting copyrighted works, especially when they reproduce or adapt substantial portions of such works. This area of law is still evolving, and specific circumstances will likely be highly determinative in any legal considerations.

23. Is the substantial similarity test adequate to address claims of infringement based on outputs from a generative AI system, or is some other standard appropriate or necessary?

The "substantial similarity" test has long been a cornerstone of copyright infringement analysis in the U.S. It determines whether an allegedly infringing work is

substantially similar to a copyrighted work in its protected elements, such that an ordinary observer would recognize the latter as having been unlawfully appropriated.

Current Adequacy

- Flexibility: The test's somewhat subjective nature offers flexibility. Given the vast range of potential outputs from AI, having a flexible standard can be beneficial.
- Precedents: Courts have a history of applying the substantial similarity test in various contexts, providing a rich jurisprudential background.
- Focus on Perception: The test's emphasis on the perspective of an "ordinary observer" can be useful when considering the vast and unpredictable outputs of AI. It's rooted in the idea that infringement is often perceptual.

Challenges with AI

- Volume and Variability of AI Outputs: Generative AI systems can produce a massive volume of outputs at rapid speeds. Determining substantial similarity in each instance might be impractical.
- Granularity of Infringement: AI might reproduce very granular elements of a copyrighted work, possibly embedded within larger unique compositions. This makes it challenging to determine when substantial similarity arises.
- AI's Learning Mechanism: AI models, especially deep learning models, function as "black boxes." Understanding how they decide to generate particular outputs can be inscrutable, complicating the infringement analysis.
- Nuanced Infringements: AI might not reproduce content verbatim but might capture its essence or structure, leading to nuanced infringements that challenge the traditional substantial similarity test.

Alternative Approaches

- Probabilistic or Quantitative Tests: Given the mathematical underpinnings of AI, introducing a more quantitative approach to determining infringement might be explored. For example, statistical measures could assess the likelihood that an AI's output was derived from a specific copyrighted source.
- Functional Analysis: A focus on the functional aspects of how AI systems produce outputs, rather than solely the outputs themselves, could provide a more comprehensive understanding of potential infringements.

● Threshold-Based Analysis: Establishing certain thresholds (e.g., percentage of similarity, recognizability of source material) could help streamline the assessment process for AI-generated outputs.

In conclusion, while the substantial similarity test offers a historically rooted and flexible approach, the unique challenges posed by AI-generated content suggest that adaptations or complementary standards might be necessary. It's essential to strike a balance between protecting copyright owners' rights and fostering innovation in the AI space. As AI continues to evolve and its applications expand, legal standards may need to be revisited and refined accordingly.

24. How can copyright owners prove the element of copying (such as by demonstrating access to a copyrighted work) if the developer of the AI model does not maintain or make available records of what training material it used? Are existing civil discovery rules sufficient to address this situation?

In copyright infringement cases, plaintiffs must prove two elements: (1) that the defendant actually copied the copyrighted work, and (2) that the copying is substantial enough to constitute an infringement. Demonstrating that a defendant had access to the copyrighted work is one way to prove the first element.

Challenges in the Context of AI:

- Volume of Data: AI models, especially those like deep learning networks, are trained on vast datasets, potentially encompassing millions of individual items. Tracking each item can be challenging.
- Opacity of Models: AI models, particularly deep learning models, are sometimes characterized as "black boxes." It can be challenging to discern precisely how they generate outputs based on inputs.

Methods to Prove Copying without Training Data Records:

- Striking Similarity: If the output of an AI system is so strikingly similar to a copyrighted work that it's unlikely the similarity arose by coincidence, it might be evidence of copying, even if direct access isn't demonstrated.

- Patterns of Outputs: If an AI system consistently generates outputs similar to a particular copyrighted work or set of works, it might suggest the model was trained on those works.
- Expert Testimony: Experts can analyze the behavior and outputs of an AI model to make informed assessments about its likely training data.
- Forensic Analysis: In some cases, it might be possible to conduct a forensic analysis of the AI system or its underlying code to discern clues about its training data.

Existing Civil Discovery Rules:

The current civil discovery rules allow parties to request documents, interrogate witnesses, and conduct depositions to uncover evidence in litigation. However:

● Data Volume: Given the massive amount of training data that can be involved with AI, it might be burdensome or even unfeasible to produce all the relevant data during discovery.
● Trade Secrets & Confidentiality Concerns: AI developers might resist discovery requests related to their training data, algorithms, and model architectures on the grounds that they constitute trade secrets or confidential business information.
● Technological Hurdles: Extracting meaningful information about training data from a trained AI model can be technologically challenging and might not always yield clear answers.

In conclusion, while proving copying in the context of AI can be challenging without records of training data, there are still methods available to copyright owners. However, given the unique challenges posed by AI, there might be a need to revisit or adapt existing discovery rules or establish new guidelines specifically tailored for cases involving AI systems.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties?

Determining liability in the context of AI-generated material can be complex because multiple entities are often involved in the creation, distribution, and use of the AI system. Here's an analysis of potential parties and their possible liabilities:

● Developer of the Generative AI Model:
○ Direct Liability: If the developer knowingly used copyrighted materials without permission as training data, they could be directly liable for infringement. Their intent and knowledge would be critical factors.
○ Secondary Liability: If the developer provides tools or models that they know or have reason to know are being used to infringe copyrights, they might face secondary liability.

● Developer of the System Incorporating the Model:
○ Direct Liability: If they are actively using the model to produce and distribute infringing content, they could be held directly liable.
○ Secondary Liability: If they provide a platform or service that they know is being used for infringement, they might be held secondarily liable. Factors such as the ability to control the infringing activity and direct financial benefit from the infringement could be considered.

● End Users of the System:
○ Direct Liability: If end users utilize the AI system to produce and distribute copyrighted content without permission, they can be held liable. This would be especially clear if users provide specific prompts or inputs to generate close replicas of copyrighted content.
○ Secondary Liability: This is less likely for end-users unless they facilitate or induce others to infringe copyright using the AI system.

● Other Parties:
○ Distributors, retailers, or platforms that host AI-generated content could potentially be held liable if they have knowledge of infringement and contribute to it. The principles of the Digital Millennium Copyright Act (DMCA) and its safe harbor provisions, which currently protect online

platforms from certain copyright liabilities, might be looked at for guidance.

● Factors Influencing Liability:
○ Knowledge and Control: Entities that have knowledge of the infringement and the ability to control it are more likely to be held liable.
○ Economic Benefit: Benefiting directly from the infringement can be a factor.
○ Nature of Infringement: Was the infringement willful or unintentional? Was it a direct replica or a transformative work?
○ Technical Measures: Did the entity implement any technical measures to prevent or mitigate infringement?

It's worth noting that the application of existing copyright doctrines to AI is still evolving. As AI technologies advance and their societal roles expand, legal standards and precedents will continue to adapt. Ultimately, determining liability will likely depend on the specifics of each case, including the roles of the various parties involved, their intent, and the degree of their involvement in the infringing activity.

25.1. Do “open-source” AI models raise unique considerations with respect to infringement based on their outputs?

Open-source AI models are publicly accessible and typically come with licenses that allow for modifications, redistribution, and sometimes even commercial uses. Because of the open and collaborative nature of these models, they do introduce some unique considerations regarding copyright infringement based on their outputs:

● Broad Accessibility and Use:
○ Since open-source models are available to a wide audience, determining responsibility for specific infringing outputs can be challenging. Many entities or individuals might use or modify the model, leading to a distributed responsibility.
● Model Modifications:

◯ The open-source nature of these models means that they can be modified and improved upon by multiple contributors. If any of these contributors introduce copyrighted material without proper permissions, it can pose a risk to subsequent users of the modified model.

● Licensing Ambiguities:
◯ While the model itself might be open-sourced, the training data used could still be proprietary or copyrighted. The licenses attached to open-source models might not always clarify the rights and restrictions concerning the training data.

◯ Some licenses may also not address the copyright implications of the outputs generated by the model, leading to uncertainties in their legal use.

● Redistribution and Derivative Works:
◯ Open-source licenses often allow redistribution. If an AI model, originally open-source but later modified with copyrighted data, gets redistributed, it can amplify the risk of infringement. Users may be unaware of the model's tainted training data and could inadvertently produce infringing outputs.

● Provenance and Accountability:
◯ Due to the collaborative nature of open-source projects, tracking the origin of specific data or modifications can be challenging. This can make it difficult to ascertain accountability if copyrighted material is found to have been used in the model's training.

● Community Vigilance:
◯ One of the advantages of open-source projects is the community's vigilance. If copyrighted material is identified within the training data or if the model generates outputs that are potentially infringing, the community might be quick to flag and rectify the issue.

In summary, while open-source AI models democratize access to advanced technology, they also introduce complexities regarding copyright implications of their outputs. Proper documentation, clear licensing, and community vigilance are crucial to mitigating potential infringement risks. Users of open-source AI models should exercise caution, ensuring they understand the model's training data and any potential copyright pitfalls associated with its outputs.

26. If a generative AI system is trained on copyrighted works containing copyright management information, how does 17 U.S.C. 1202(b) apply to the treatment of that information in outputs of the system?

17 U.S.C. 1202(b) is a provision of the Digital Millennium Copyright Act (DMCA) that addresses the unauthorized removal or alteration of copyright management information (CMI). Specifically, 17 U.S.C. 1202(b) states:

"No person shall, without the authority of the copyright owner or the law— (1) intentionally remove or alter any copyright management information, (2) distribute or import for distribution copyright management information knowing that the copyright management information has been removed or altered without authority of the copyright owner or the law, or (3) distribute, import for distribution, or publicly perform works, copies of works, or phonorecords, knowing that copyright management information has been removed or altered without authority of the copyright owner or the law, knowing, or, with respect to civil remedies under section 1203, having reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement of any right under this title."

For generative AI systems, several considerations emerge when thinking about this provision:

● AI Outputs and CMI: If a generative AI system creates new content that is entirely its own and not a direct reproduction of the input, it might not necessarily carry over the CMI from the training data. However, if it does reproduce parts of the copyrighted work verbatim, and in the process, omits or alters the CMI, there might be a concern under 1202(b).
● Intentionality: The statute mentions "intentionally remove or alter." Given that AI models process data in ways that don't inherently exhibit human-like intent, it might be challenging to argue that an AI system "intentionally" removed

CMI. However, developers, operators, or users who knowingly use AI to bypass CMI might be subject to scrutiny.

● Knowledge Requirement: The provision requires a party to "know" that CMI has been removed or altered without authority. Given that generative AI might produce a vast array of outputs, it may be challenging for users or developers to be aware of each instance where CMI is affected.

● Embedding CMI in AI Outputs: An interesting potential solution could be the development of AI models that can recognize and preserve or even embed CMI into their outputs. Such a feature would provide a proactive approach to respecting copyrights.

● Practical Application: Given the abstract and high-dimensional nature of how AI models process data, directly linking the removal of CMI in the model's internal operations to the language of 17 U.S.C. 1202(b) might be challenging from a legal standpoint.

In conclusion, the application of 17 U.S.C. 1202(b) to generative AI systems trained on copyrighted works with CMI remains an evolving area of law. As AI continues to advance and its use becomes more widespread, there might be a need for clearer guidelines or legislative updates to address these nuances.

27. Please describe any other issues that you believe policymakers should consider with respect to potential copyright liability based on AI-generated output.

The intersection of AI and copyright is a complex domain with evolving implications. Policymakers should consider several additional issues when thinking about potential copyright liability for AI-generated output:

● Moral Rights: While the U.S. has limited recognition of moral rights, other jurisdictions emphasize the personal and reputational aspects of an author's connection to their work. How AI-generated outputs might impact these rights, especially when works are altered or presented in new contexts, needs consideration.

● Attribution and Authenticity: There's a potential risk of AI-generated content being mistaken for human-created content. Clarifying standards for attribution

can help audiences differentiate between AI and human outputs and prevent misrepresentation.

● Evolving Nature of AI: AI models continually improve, meaning today's concerns might differ from those in a few years. Policymakers should adopt a forward-looking perspective, building flexibility into any legal frameworks. Global Implications: AI operates across borders. Harmonizing copyright standards or at least understanding international variations is crucial for developers and businesses operating in multiple jurisdictions.

● Economic Impacts: Overly restrictive copyright policies could stifle innovation and the development of beneficial AI technologies. Conversely, too lax an approach might undermine content creators' rights and economic incentives. Fair Dealing and Parody: How AI-generated outputs that could be considered parodies or satirical commentaries fit into the copyright landscape needs clarity. Parody is often an exception to copyright infringement in many jurisdictions.

● Archival and Research Concerns: Some AI research involves massive datasets that could include copyrighted materials. Clear guidelines are needed to ensure that academic and institutional research can proceed without undue hindrance while respecting copyright.

● Potential for Automated Licensing: As AI systems become more sophisticated, there's potential for automated licensing mechanisms. For instance, AI could negotiate licensing agreements or determine appropriate royalty payments based on usage, potentially streamlining rights management.

● Public Domain and Open Source: How AI interacts with public domain materials and open-source licenses could warrant separate considerations. Using these materials for training might be more permissible, but ensuring that AI-generated outputs don't inadvertently infringe upon other works is essential.

● Interdisciplinary Collaboration: The complexity of AI and its intertwining with copyright means legal experts alone might not capture the full picture. Collaborations between technologists, ethicists, artists, and legal experts could yield more comprehensive and balanced policies.

● Educational and Awareness Campaigns: As AI becomes more entrenched in daily life, there's a need to educate both creators and the general public about the copyright implications of AI-generated outputs.

In summary, the potential copyright liability based on AI-generated outputs is an intricate and multi-faceted issue. A balanced, informed approach that protects creators' rights while promoting technological innovation is essential as we navigate this new frontier.

28. Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

The question of labeling AI-generated material is a nuanced one with potential implications for transparency, consumer protection, and fairness. Here are some arguments for and against such a requirement, followed by considerations for its implementation:

Arguments in Favor of Labeling AI-Generated Material:

● Transparency: It promotes honesty and openness about the origins of content, helping users understand the source and potential biases of the information they encounter.
● Consumer Protection: AI-generated content might be used in deceptive ways, such as fake reviews or AI-generated endorsements. Labeling can help prevent consumer deception.
● Ethical Considerations: It respects the right of individuals to know whether they're interacting with or viewing content produced by a machine rather than a human, especially in sensitive contexts like therapy or news dissemination.
● Creative Integrity: It helps differentiate human creativity from machine-generated content, preserving the unique value of human authorship.

Arguments Against Labeling AI-Generated Material:

● Stifling Innovation: Mandatory labeling might discourage the development and adoption of beneficial AI technologies in creative fields.
● Practicality: Given the sheer volume of AI-generated content and the various ways AI is employed, labeling every piece might be impractical.
● Potential Stigmatization: There might be an undue bias against AI-generated content, leading to its automatic devaluation even if it's of high quality or utility.

Implementation Considerations:

● Scope: Determine in which contexts labeling is most crucial. Is it more important for news articles and academic papers than for entertainment or art?
● Labeling Mechanism: How should AI-generated content be labeled? Visual markers, metadata tags, or other identifiers could be used.
● Granularity: Should the label specify the role of AI? For instance, differentiate between AI-generated, AI-assisted, or AI-curated content.
● Exceptions: There might be situations where labeling isn't necessary or where exceptions should be made, such as experimental art projects.
● Enforcement: Establish a mechanism to monitor compliance and address instances where AI-generated content isn't appropriately labeled.
● Education: Alongside labeling, there's a need for public education on what AI-generated content means and its implications.
● Global Consistency: Since AI and digital content often transcend borders, international standards or agreements might be beneficial.

In summary, while labeling AI-generated material has its merits, especially in the interest of transparency and consumer protection, it also comes with challenges. A one-size-fits-all approach might not be suitable, and a nuanced, context-dependent framework may be more appropriate.

28.1. Who should be responsible for identifying a work as AI-generated?

Determining responsibility for identifying a work as AI-generated can be a complex issue, but here are some potential parties that could bear this responsibility, depending on the context:

● Developers of the AI System: Those who create and sell or distribute AI software or platforms might be best positioned to implement identification features, especially if the AI is intended for broad public use. For example, if an AI tool is designed to produce music, the developer could embed a mechanism that tags every piece it generates.
● Operators or Users of the AI System: People or entities that deploy AI systems for specific tasks, whether it's writing articles, creating artwork, or generating videos, should label such content as AI-generated. They are the

ones who choose to publish or release the content and, as such, hold the primary responsibility for its labeling.
● Distributors or Platforms: Online platforms, such as social media sites, content distributors, or digital marketplaces, could establish guidelines or requirements for users to label AI-generated content. This can be likened to platforms today that label certain content as "sponsored" or "promoted."
● Content Curators or Editors: In contexts where AI-generated content is mixed with human-generated content, such as news websites or journals, the curators or editors might be responsible for ensuring that AI contributions are labeled appropriately.
● Regulatory or Oversight Bodies: In certain industries or regions, regulatory bodies might mandate the labeling of AI-generated content to ensure consumer protection, transparency, or ethical considerations. These bodies would be responsible for creating standards and ensuring adherence.
● AI System Itself: In some cases, the AI system could be designed to automatically label its outputs as AI-generated, ensuring that any content it produces, regardless of where it's used, is appropriately identified.

The responsibility might vary based on the specific use-case, the intent of the AI's deployment, the potential impact on end-users, and the jurisdiction in which the AI operates. It's also possible that a combination of the parties listed above would share this responsibility to ensure broad compliance and awareness.

28.2. Are there technical or practical barriers to labeling or identification requirements?

Yes, there are several technical and practical barriers to labeling or identification requirements for AI-generated content:

● Scale of Content Generation: The sheer volume of content that can be generated by AI, especially in automated settings, can make consistent labeling challenging. AI can produce content at a pace much faster than humans can monitor.
● Embedding in Different Media Types: Depending on the type of content, embedding a label can be technically challenging. While it's straightforward to add a label to a piece of text, doing so for images, music, or videos might require more intricate methods.

● Manipulation After Generation: Once AI-generated content is produced and disseminated, third parties can easily modify, repurpose, or redistribute it, potentially removing any AI labels in the process.
● Decentralized and Open-Source AI Models: Many AI models, especially those that are open-source, can be used by a wide array of individuals and entities across different jurisdictions. Enforcing consistent labeling across such a decentralized ecosystem is challenging.
● Legacy Content: If labeling becomes a new standard or requirement, addressing pre-existing AI-generated content that hasn't been labeled poses a practical challenge.
● Ambiguity in "AI-Generated" Definition: Determining what qualifies as "AI-generated" can be ambiguous. For instance, if a human edits or refines AI-generated content, at what point does it cease to be "AI-generated"?
● Resistance from Stakeholders: Some creators or companies might resist labeling requirements, fearing it could devalue AI-generated content or create biases against it.
● International Variability: Different countries might have different regulations or standards for labeling AI-generated content. This can be a challenge for platforms or entities operating globally.
● Detection and Enforcement Challenges: Monitoring and enforcing labeling requirements across the vast digital landscape is not trivial. It would require significant resources and potentially new technologies to detect unlabeled AI-generated content.
● Overhead and Performance Concerns: In real-time systems, where AI generates content on-the-fly (like in video games or interactive applications), adding a labeling process might introduce performance overhead.

While these challenges are significant, they aren't insurmountable. With coordinated efforts, clear standards, technological advancements, and stakeholder buy-in, effective labeling mechanisms can be developed and implemented.

28.3. If a notification or labeling requirement is adopted, what should be the consequences of the failure to label a particular work or the removal of a label?

If a notification or labeling requirement for AI-generated material is adopted, the consequences for failing to label a work or for removing a label could vary based on

the severity of the offense, intent, and the potential harm caused. Here are some potential consequences that could be considered:

● Warnings: Initial failures could result in warnings, allowing the responsible party to correct the oversight. This approach recognizes that errors can happen and provides a chance for rectification.
● Fines or Penalties: Repeat violations or intentional removal of labels could lead to monetary penalties. The amount could be based on the severity of the offense or be graduated based on repeated offenses.
● Liability for Misrepresentation: If unlabeled AI-generated content leads to any form of harm, misrepresentation, or deception, the responsible party could be held liable.
● Suspension or Ban: Platforms could temporarily suspend or permanently ban users or entities that repeatedly fail to label AI-generated content or intentionally remove labels.
● Revocation of Licenses: If the entity operates under a specific license (for instance, a broadcasting license), repeated violations could lead to the license being revoked.
● Take down Requests: Platforms could be required to take down unlabeled AI-generated content once detected or reported.
● Mandatory Corrective Actions: In cases where AI-generated content without labels causes confusion or misinformation, entities could be required to issue corrections or clarifications.
● Restitution: If the failure to label results in financial or reputational harm to others, the party responsible for the unlabeled content might be required to compensate the affected parties.
● Criminal Charges: In extreme cases, especially where the unlabeled content leads to significant harm, fraud, or deception, criminal charges could be considered.
● Enhanced Scrutiny: Entities with repeated labeling violations might be subjected to increased scrutiny or audits.
● Public Disclosure: A registry or database could be established where repeat offenders are listed, making it public knowledge and potentially acting as a deterrent.

● Reputation Impact: Intentional mislabeling or removal of labels might tarnish the reputation of an entity, especially if they are in the business of creating or distributing content.

In establishing consequences, it's important to consider a balanced approach that deters non-compliance but doesn't stifle innovation or unnecessarily burden creators. Factors like intent (accidental vs. intentional), the potential harm caused by the unlabeled content, and the track record of the offending party should be taken into account.

29. What tools exist or are in development to identify AI-generated material, including by standard-setting bodies? How accurate are these tools? What are their limitations?

The identification of AI-generated material is a rapidly evolving field as generative models continue to improve and produce increasingly realistic outputs. Several tools and techniques have been developed or are in development to distinguish between human-generated and AI-generated content. Some of the prominent ones include:

● Deep Learning Forensics: These are models designed to detect the subtle patterns, artifacts, or inconsistencies in AI-generated content. For example, in the domain of deepfake videos, specialized neural networks can identify inconsistencies in lighting, blinking patterns, or artifacts that humans might overlook.
● Metadata Analysis: AI-generated content might leave traces in metadata that can reveal its origin. By analyzing this metadata, one can identify the tools or platforms used to generate the content.
● Reverse Image/Video Search: Tools like Google's reverse image search can be used to identify original images or videos that might have been manipulated or used as a base for AI-generated content.
● Consistency Checks: AI-generated text, especially when produced at length, can sometimes have inconsistencies in narrative or details. Tools can be developed to analyze and flag such inconsistencies.
● Steganography: Watermarking and other steganographic techniques can be used to embed information into content, which can then be used to verify the authenticity of a piece of content.

● Blockchain Verification: Blockchain can be used to store digital signatures of authentic content. Any piece of content can then be verified against this immutable record to determine its authenticity.
● Physical World Verification: For AI-generated images or videos, inconsistencies with the physical world (like reflections, shadows, or the physics of moving objects) can be identified.
● Audio Analysis: For AI-generated audio or voice deepfakes, spectral analysis and other audio forensic techniques can be used to detect inconsistencies or artifacts.
● Standard-Setting Bodies: Bodies like the International Standards Organization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) are actively involved in setting standards for AI. They may develop benchmarks and best practices for the detection and identification of AI-generated content.

Limitations:

● Evolving Models: As generative models like GANs (Generative Adversarial Networks) improve, the distinctions between real and AI-generated content become finer, challenging the capabilities of detection tools.
● False Positives/Negatives: No tool is perfect, and there's always a risk of false positives (labeling real content as AI-generated) and false negatives (failing to detect AI-generated content).
● Computational Costs: Some detection methods can be computationally intensive, making them less practical for real-time or large-scale use.
● Generalization: A tool trained to detect one kind of AI-generated content might not perform well on content generated by a different or newer model.

Accuracy:

● The accuracy of these tools can vary widely based on their design, the domain (text, image, audio, video), and the sophistication of the AI-generating model. Generally, as AI models evolve, there's a kind of "arms race" between generative models and detection tools.

Overall, while tools for detecting AI-generated content are improving, they need constant updates to keep pace with advancements in generative AI technologies. Collaboration between researchers, industries, and policymakers can help ensure that these tools remain effective.

30. What legal rights, if any, currently apply to AI-generated material that features the name or likeness, including vocal likeness, of a particular person?

The legal rights surrounding AI-generated material that features the name or likeness of a particular person largely fall outside the realm of copyright law and instead intersect with areas such as rights of publicity, defamation, and privacy law. These rights and laws can vary significantly depending on the jurisdiction, but here's a broad overview of the U.S. context:

- Rights of Publicity: This refers to an individual's right to control the commercial use of their name, image, voice, or other elements of their identity. If an AI-generated material (like a deepfake video or synthetic voice) commercially exploits a person's identity without consent, it could infringe upon their right of publicity. The specifics of these rights vary by state, with some states, like California, having robust statutory protections while others might offer limited or no formalized protection.
- Defamation: If AI-generated content portrays an individual in a false light that harms their reputation, it could give rise to a defamation claim. This would be true if, for instance, a deepfake video falsely depicted someone engaging in illegal or immoral behavior.
- Invasion of Privacy: Depending on how AI-generated content uses an individual's likeness or personal information, it might lead to claims of invasion of privacy. These claims can arise in various contexts, such as the public disclosure of private facts or the portrayal of someone in a false light.
- Unfair Competition & False Endorsement: Under the Lanham Act, if an AI-generated representation of a person is used in a way that falsely suggests that the individual endorses a product or service, it might give rise to claims of false endorsement or unfair competition.
- Consent & Contractual Rights: Many celebrities and public figures have contracts or provisions that control the use of their name, image, or voice. Unauthorized use in AI-generated material could breach these agreements.
- First Amendment Considerations: In the U.S., the First Amendment protects freedom of speech and expression. Courts will need to balance the rights of creators of AI-generated content with the rights of individuals whose likenesses are used. Creative, transformative, or newsworthy uses of an

individual's likeness may receive First Amendment protections against certain claims.

It's essential to recognize that the rapid advancement of AI and related technologies is continually challenging the boundaries of these legal concepts. Courts and legislatures are grappling with how to apply traditional legal frameworks to these novel situations. As a result, the legal landscape in this area is dynamic and will likely evolve in response to technological changes and high-profile cases.

31. Should Congress establish a new federal right, similar to state law rights of publicity, that would apply to AI-generated material? If so, should it preempt state laws or set a ceiling or floor for state law protections? What should be the contours of such a right?

The question of whether Congress should establish a federal right of publicity, specifically to address AI-generated material, is a complex issue involving multiple considerations. Here are some arguments for and against, along with issues to consider:

Arguments for a Federal Right:

- Uniformity: A federal statute could provide a uniform set of rules that would make it easier for creators, users, and affected individuals to understand their rights and obligations. This could reduce the legal complexity that currently arises from varying state laws.
- Clarity for AI Developers: A federal law could establish clear guidelines for AI developers regarding what uses of a person's likeness would be permissible, thus promoting innovation and reducing the risk of litigation.
- Consumer Protection: A federal statute could ensure that individuals nationwide have at least a minimum level of protection against unauthorized exploitation of their name, image, or likeness in AI-generated content.
- Adaptability: Federal legislation could be crafted to address the unique challenges posed by AI, such as deepfakes or voice synthesis, that may not be adequately addressed by existing state laws.

Arguments Against a Federal Right:

- Federalism Concerns: Rights of publicity have traditionally been governed at the state level. A federal statute could override long-standing state laws and traditions, undermining the ability of states to regulate based on their unique circumstances.
- Overreach and Unintended Consequences: There is a risk that a poorly drafted federal statute could overly restrict freedom of expression, stifle creativity, or hinder technological innovation.
- Complexity: Creating a federal framework could add an additional layer of complexity to an already complicated legal landscape, particularly if it doesn't preempt state laws.

Factors to Consider:

● Scope: The law would need to define what "likeness" means in the age of AI, considering voices, avatars, and even behavioral traits that could be mimicked by AI.
● Exemptions: Any law would need to care fully balance commercial exploitation with free expression, possibly allowing exemptions for news reporting, satire, or research.
● Enforcement: The law should specify who can bring a claim, under what circumstances, and what damages or remedies are available.
● State Law Interaction: Congress would need to decide whether a federal law would preempt state laws or provide a floor/ceiling for state law protections. Each approach has its own advantages and disadvantages.
● Technological Neutrality: Given the rapid pace of technological change, it would be advisable for the legislation to be technologically neutral, allowing it to adapt to future developments.

Creating a new federal right would be a significant legislative undertaking that would need to be done with considerable care to balance the interests of individuals, creators, and the tech industry.

32. Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator (such as an AI system producing visual works “in the style of” a specific artist)? Who should be eligible for such protection? What form should it take?

The question of whether AI-generated content that imitates the artistic style of a human creator should be protected or regulated touches on a delicate balance between encouraging innovation and respecting the rights of artists. Here's a breakdown of the considerations:

Arguments for Protections:

- Protecting Unique Artistic Expression: An artist's style can be seen as a culmination of their experiences, skills, and emotions. If an AI can easily replicate and mass-produce this style, it might diminish the value and uniqueness of the artist's work.
- Economic Concerns: An artist might lose potential revenue if AI-generated art in their style saturates the market or is sold without their permission.
- Misattribution and Reputation: There's a risk of misattribution, where consumers might believe that the AI-generated content was created by the human artist. This could harm the artist's reputation, especially if the quality of the AI-generated work is subpar.

Arguments Against Protections:

- Stifling Innovation: Overly strict regulations might stifle advancements in AI and limit its potential benefits for artistic creation, collaboration, and more.
- Blurry Lines: Artistic styles are often influenced by other artists, historical movements, and common techniques. Defining what constitutes a unique "style" might be too ambiguous for legal standards.
- Freedom of Expression: Limiting the generation of art, even by AI, can be seen as a restriction on freedom of expression.

Potential Forms of Protection:

- Right of Attribution: Artists could have the right to be attributed (or not attributed) when an AI generates work in their style.
- Licensing and Royalties: There could be a system where artists can license their style for AI usage, earning royalties when it's used.
- Opt-Out System: Similar to the "right to be forgotten" in some privacy laws, artists could request that specific AI systems avoid replicating their style.

- Limitations on Commercial Use: AI-generated art that mimics a specific artist's style might be allowed for personal or educational uses but restricted for commercial purposes.

Eligibility:

● Established Artists: Protections might only apply to artists who've reached a certain level of recognition or can demonstrate the uniqueness of their style.
● Duration: Like copyright, which lasts for a set number of years, protection for an artist's style might be time-limited.
● Register or Apply: To avoid over burdening the system, artists might need to register or apply for such protection proactively.

In conclusion, while it's essential to protect artists' rights and contributions, any legal approach must be carefully crafted to ensure it doesn't stifle innovation or overly restrict the development and use of AI in the arts.

33. With respect to sound recordings, how does section 114(b) of the Copyright Act relate to state law, such as state right of publicity laws? Does this issue require legislative attention in the context of generative AI?

Section 114(b) of the Copyright Act addresses the scope of exclusive rights for sound recordings. Specifically, it clarifies that the exclusive rights of a copyright owner in a sound recording are limited to the rights specified in section 106(6), and those rights apply only to activities that directly involve the use of the copyrighted sound recording. The section is designed to emphasize that the copyright in a sound recording doesn't extend to the underlying musical or literary work that the recording embodies.

In essence, Section 114(b) ensures that a sound recording copyright holder's rights are confined to the specific recording, and not the underlying musical composition or performance. For example, just because someone holds the copyright to a particular recording of a Beethoven symphony doesn't mean they hold rights to all recordings or performances of that symphony.

Now, when we bring state rights of publicity into the picture, things get more complicated. Rights of publicity generally protect individuals (often celebrities or

public figures) from unauthorized commercial exploitation of their name, image, voice, or likeness.

In the context of sound recordings, there could be a tension between the federal copyright protections in Section 114(b) and state-level rights of publicity. For instance, if an AI system were to generate a song that sounds like it's sung by a famous artist without that artist's permission, even if it doesn't infringe on the sound recording copyright, it might violate the artist's right of publicity under state law if it suggests an endorsement or affiliation that isn't real.

Legislative attention might be required in the context of generative AI for several reasons:

● Clarification: It might become necessary to clarify how Section114(b) applies to AI-generated sound recordings, especially if those recordings resemble or mimic the style or voice of established artists.
● Uniformity: Given the variability of rights of publicity laws across states, a federal standard or guideline might be beneficial for AI developers and users to have a consistent understanding of their boundaries and liabilities.
● Evolution of Technology: Generative AI technologies are rapidly evolving, and their capabilities in generating sound recordings that closely resemble existing works or artists' styles are improving. As these technologies progress, it will be essential to reassess and potentially adjust the legal framework to reflect these advancements.

In conclusion, while Section 114(b) and state rights of publicity laws address different aspects of sound recordings and artist rights, the rise of AI and its capabilities in this realm might necessitate legislative attention to ensure both clarity and fairness in the law.

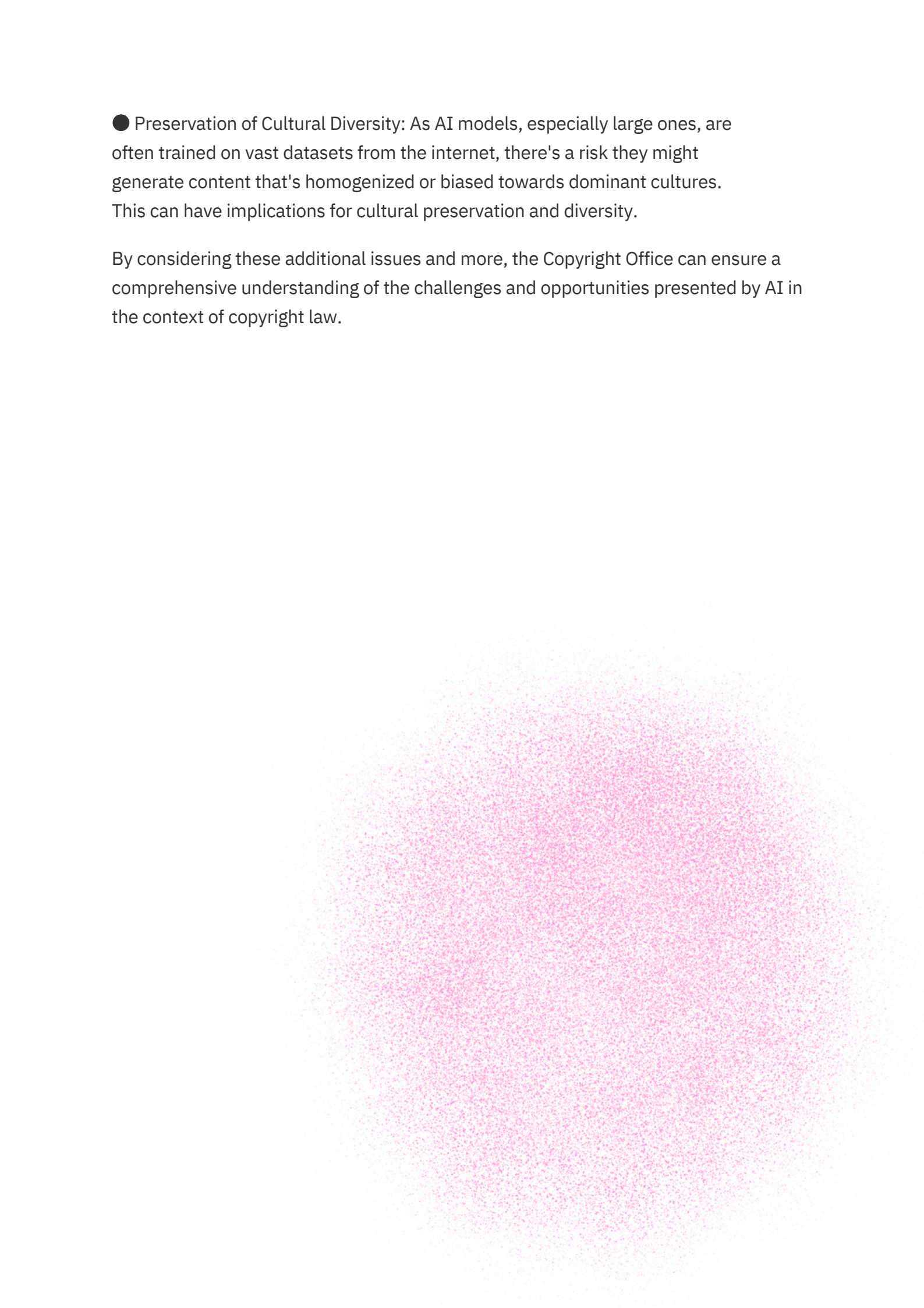
34. Please identify any issues not mentioned above that the Copyright Office should consider in conducting this study.

The introduction of AI into the world of creativity and copyright brings forth a variety of challenges and considerations that require close examination. Beyond the issues already mentioned, the Copyright Office might consider the following:

- Ethical Implications: The potential for AI to create derivative works or new creations raises significant ethical questions. For example, how should AI-generated works be approached when they deal with culturally sensitive materials or topics?
- Archival and Historical Integrity: If AI begins producing art, music, literature, and more, how do institutions archive and document these creations? What measures need to be in place to ensure that future generations understand the context in which these AI-generated materials were created?
- International Considerations: How are other countries addressing AI in the realm of copyright? Are there international standards or agreements that the U.S. should be aware of or consider aligning with?
- Technological Evolution and Future-Proofing: Given the rapid pace of technological change, any legislative or regulatory changes should consider how to remain relevant and adaptable to future advancements in AI and machine learning.
- AI as a Tool vs. AI as a Creator: There's a distinction between AI being used as a tool by a human creator and AI autonomously generating content. This difference might warrant varied levels of copyright protection or recognition. Interplay with Other Intellectual Property Rights: How does the use of AI intersect with other areas of intellectual property, such as patents or trademarks? For instance, could AI-generated logos or brand names be trademarked?
- Economic Impacts: How might the rise of AI-generated content affect various industries, from publishing and music to film and gaming? Consideration should be given to the economic repercussions for individual creators and industries at large.
- Authorship Attribution: How do we attribute authorship for AI-created works? If an AI creates a piece of music that becomes a hit, who gets credited as the "artist"?
- Education and Public Awareness: As AI becomes more integrated into the world of content creation, there may be a need for public education campaigns to help people understand the implications and nuances of AI-generated content.
- Privacy Concerns: If AI systems are trained on personal or private data, there could be concerns about potential leaks or misuse of such data, especially if it inadvertently appears in generated content.

● Preservation of Cultural Diversity: As AI models, especially large ones, are often trained on vast datasets from the internet, there's a risk they might generate content that's homogenized or biased towards dominant cultures. This can have implications for cultural preservation and diversity.

By considering these additional issues and more, the Copyright Office can ensure a comprehensive understanding of the challenges and opportunities presented by AI in the context of copyright law.



SeaQVN Info: nitisharora@seaqvn.com | Phone: +91-9953831246 | Address: BH308, 81 Business Hub, Sector 81, Greater Faridabad, Faridabad, Haryana, India PIN:121007