Who Owns the Data Behind Your AI?

Melissa Heidrick

Jan 01, 2025 11 min read

Summary

- Understand the legal and ethical implications of AI training data provenance.
- Explore key court cases shaping intellectual property law.
- Learn actionable strategies for attorneys to counsel clients, audit AI tools, and lead in an AI-dominant landscape.

iStock.com/quantic

Could the way your generative artificial intelligence (GenAI) tools are trained expose you or your clients to infringement claims? If this question unsettles you, you're not alone. The popularity of AI tools continues to increase, to the point that even regulators and bar associations are providing guidance on navigating the associated legal and ethical

challenges. And, while many of us are eagerly embracing the benefits of these tools, their adoption has also sparked a wave of intellectual property questions and disputes over pretraining data. These debates highlight questions about ownership, authorship and ethical use, and could impact both AI liability and intellectual property frameworks––making it critical for legal professionals to pay attention.

Before GenAI tools reach end users, they are pretrained on large datasets. But where does the training data behind your AI come from? And does its use infringe on the rights of others?

Improperly sourced training data could lead to copyright infringement claims, privacy violations or ethical breaches. Additionally, secondary liability is becoming an increasingly prominent issue. Courts and regulators are scrutinizing the platforms, developers and corporate officers involved in creating and deploying these tools. With the rapid inclusion of GenAI features in legal technology tools, this expanding scope highlights questions about contributory infringement, vicarious liability and corporate accountability.

As the legal landscape surrounding these issues evolves, identifying and proactively addressing these risks is critical. Attorneys must grapple with how to counsel clients using or building AI tools while ensuring their own AI use remains above board.

Pretraining: The P in GPT

Before GenAI models reach end users, they must undergo pretraining––or, the “P” in GPT generative pretrained processor). Pretraining provides the data-based knowledge needed for the model to generate human-like text, create images or perform other sophisticated tasks.

So, how does pretraining work? It involves collecting and providing data to an AI model. The large nature of these datasets enables a model to identify patterns, define context and predict favorable outputs. Analyzing billions of textual examples from books, articles, websites and anything else placed in the training set enables large language models to identify how words, phrases and sentences relate to one another.

- **Tokenization.** Before training begins, the raw data is broken down into smaller components called tokens. For text, this can mean splitting sentences into words or “subwords,” while for images, it can involve pixel-level information.

- **Pattern recognition.** The model identifies patterns and relationships within these tokens. For example, it might identify that "contract" is often found near the word "terms" or that repetitive clauses frequently appear in legal documents. (E-

discovery practitioners will recognize this identification of words and their proximity to each other to identify and define concepts as the same methodology used to create proximity search terms, conceptual clustering analytics and continuous active learning workflows.)

- **Contextual understanding.** Using techniques like attention mechanisms, the model determines which parts of the input data are most relevant to the context. This is crucial for generating coherent and contextually meaningful outputs.

- **Prediction and adjustment.** The model is tasked with predicting the next token in a sequence (e.g., what word follows "party of the first part" in a legal document). If its predictions are incorrect, the model adjusts its internal parameters through a process called backpropagation, improving its accuracy over time.

- **Iteration over scale.** This process repeats billions of times, resulting in a model that can generate outputs that seem intelligent and relevant to the task requested by the end user.

Provenance: The Data Behind Pretraining

Training data provenance refers to the origin of the datasets used in pretraining. In other words, where did the data come from, and was its use authorized? For attorneys, this issue strikes at the heart of their obligations to practice law ethically and competently.

What makes pretraining so powerful is also what makes it precarious: the nature of the data itself. Collecting the large datasets needed for pretraining has largely taken place within a legal gray area, with huge quantities of data being scraped from publicly available sources without clear licensing or consent.

Models are often trained on publicly available and acquired datasets that can include:

- **Public text.** Online articles and blogs, discussions from chat forums, public legal filings and more.

- **Scraped data.** Information pulled from websites, often without explicit consent or licensing.

- **Proprietary content.** Licensed datasets, such as academic journals or subscription-based databases.

The scale of this data allows models to generalize across topics, but it also opens the door to serious intellectual property and ethical concerns. Though AI developers do sometimes

acquire datasets through licensing agreements with publishers or data aggregators, licensing can be prohibitively expensive, leading many to opt for scraping, instead.

Scraped datasets can include copyrighted works, proprietary business information, sensitive personal data or even data from minors. So, what happens when the training data includes copyrighted materials or sensitive personal information? Who is accountable if the outputs of the model closely mimic a protected work?

Why AI Training Provenance Matters

When it comes to training data, ignorance isn't a defense—it's a liability. Training datasets that are unlicensed, sensitive or ethically questionable pose risks for everyone involved in the AI ecosystem, affecting both legal compliance and ethical accountability.

Legal Risks

When training datasets include copyrighted or unlicensed material, they risk producing outputs that infringe on intellectual property (IP) rights. This exposure affects all participants in the AI ecosystem:

- **Developers.** Lawsuits may target developers for using protected content without authorization, raising claims of copyright infringement or trade secret violations.
- **Businesses.** Organizations deploying AI tools could face legal exposure if AI-generated outputs resemble copyrighted works. For instance, marketing teams using AI-generated content might unknowingly reproduce protected designs, triggering IP disputes.
- **End users.** Individuals using AI tools risk incorporating infringing material into their projects, leaving themselves and their firms vulnerable to legal challenges.

Ethical Concerns

Unregulated use of personal or sensitive data in training datasets raises ethical dilemmas.

- **Transparency and consent.** Was the data used to train AI collected transparently and with appropriate permissions? Without consent, the outputs could violate privacy rights or perpetuate harm.
- **Sensitive information.** AI systems trained on scraped medical information or social media posts risk misusing sensitive data, creating potential harm for

individuals. This harm becomes even more pronounced when data from minors is used.

- **Bias and representation.** Training on biased datasets can lead to models that replicate harmful stereotypes, disproportionately impacting marginalized groups.

These concerns have real-world implications that attorneys must address when using, recommending or providing counsel surrounding AI technologies.

- **Risk assessment.** The datasets used to train AI tools should be evaluated, identifying potential risks such as copyright infringement, privacy violations or unlicensed use of proprietary content. Understanding the origin and scope of the data is critical to mitigating liability.

- **Compliance strategies.** Ensuring that AI tools and their outputs comply with applicable copyright, privacy and data protection laws is key. This includes verifying proper licensing agreements and ensuring adherence to privacy regulations like the GDPR or CCPA.

- **Proactive guidance.** Attorneys should provide clients with actionable strategies for integrating AI into their operations while minimizing exposure. This includes advising on best practices for sourcing tools and anticipating potential disputes related to AI-generated outputs.

The IP Disputes Shaping the AI Landscape

Disputes over training data are raising pressing questions that could redefine intellectual property and liability frameworks. Does scraping copyrighted material for training purposes constitute infringement, or does it fall under the fair use exception? How can we ensure that training data is legally and ethically sourced? And what happens when AI outputs closely mimic protected works?

Disputes to Watch

Several high-profile cases are beginning to define the boundaries of these issues:

- ***New York Times v. Microsoft/OpenAI.*** This case alleges the unauthorized copying of millions of articles to train large language models. The *New York Times* claims both the training process and AI-generated outputs infringe on its copyrights and challenges the fundamental methods used in large-scale AI training.

- ***Authors Guild v. OpenAI (Consolidated Action).*** Prominent authors, including George R. R. Martin and John Grisham, claim OpenAI used their works without authorization for training purposes.

- ***Thomson Reuters v. Ross Intelligence.*** Thomson Reuters alleges its Westlaw headnotes and summaries were improperly used by Ross Intelligence to develop a competing AI tool.

- ***Getty Images v. Stability AI.*** Getty Images accuses Stability AI of using millions of its copyrighted photographs without authorization to train its image-generation models. This lawsuit also includes claims for trademark violations and challenges related to the removal of copyright management information such as watermarks.

- ***Andersen v. Stability AI.*** Visual artists, led by Sarah Andersen, claim that Stability AI, Midjourney and DeviantArt tools were trained on their copyrighted works without consent. The plaintiffs argue the models generate outputs that mimic their unique styles, infringing on their rights. Notably, the case recently survived a motion to dismiss key copyright claims.

- ***Universal Music Group v. Anthropic.*** Universal Music Group alleges that Anthropic's AI models reproduced copyrighted song lyrics without authorization or licensing.

- ***Vacker v. ElevenLabs (Voice Actor Cases).*** Voice actors are suing companies like ElevenLabs over the unauthorized reproduction of their voices. These cases combine claims of copyright infringement and publicity rights and could set new standards for compliance in the development of voice-based AI technologies.

The outcomes of these cases may have far-reaching implications, influencing the evolution of IP law and defining how compliance and liability are assigned across the entire AI ecosystem.

Understanding Fair Use

Fair use is the legal doctrine that permits limited use of copyrighted material without permission under certain circumstances, primarily for purposes like commentary, criticism, education or research. Courts assess fair use based on four factors.

1. **Purpose and character of the use.** Is the use transformative? Does it add new expression or meaning—or is it merely a reproduction of the original?
2. **Nature of the copyrighted work.** Creative works are afforded stronger protection than factual works.
3. **Amount and substantiality of the portion used.** How much of the original work was used, and was it the "heart" of the material?
4. **Effect on the market.** Does the use harm the market value of the original work or its potential licensing opportunities?

Developers may argue that using copyrighted material for AI pretraining qualifies as transformative. But whether training transforms the data enough to meet the standard is up for debate.

Looking Ahead: The Future of Pretraining

Alarmingly, we are running out of authentic human-generated data to use during pretraining. The availability of high-quality training data is becoming a bottleneck for AI development. Early models thrived on the abundance of freely accessible internet data, but the pool of usable training data is shrinking. Not only can we not generate it quickly enough, but awareness around intellectual property and privacy issues is growing.

Companies like OpenAI are now paying for licensed datasets from publishers—a significant shift that highlights the increasing difficulty of acquiring training data legally and ethically. So, what happens when the well of high-quality, original data runs dry? The answer could fundamentally change the trajectory of AI development—and not for the better.

The Threat of Model Collapse

As the pool of authentic data diminishes, many AI developers are turning to synthetic data—data generated by other AI models—to fill the gap. While this approach can work for a time, it introduces a significant risk: model collapse.

Unlike training on original, diverse datasets, synthetic data can create a feedback loop of diminishing quality, as errors and biases compound over successive generations. Without access to fresh, reliable data, the robustness and reliability of AI systems could erode, undermining their utility.

A Potential Solution: Monte Carlo Tree Search

To address the challenges posed by dwindling authentic training data, researchers are exploring techniques like Monte Carlo tree search (MCTS). MCTS, widely used in game theory and decision making, is a heuristic algorithm that searches for the best outcomes by simulating various decision paths and evaluating their potential results. This approach can introduce variability into AI training without relying on external datasets.

How MCTS Works

Unlike traditional Monte Carlo simulations that rely on randomness to model outcomes, MCTS builds a "tree" of possible decisions and outcomes. Each node in the tree represents a possible state, while branches represent decisions leading to other states. The algorithm iteratively simulates these paths, updates the tree with the most promising outcomes and refines its decision making over time. Key steps include:

- **Selection.** The algorithm selects the most promising node to explore based on a balance of exploration (testing new paths) and exploitation (optimizing known good paths).
- **Expansion.** The selected node is expanded by adding new potential decisions or outcomes to the tree.
- **Simulation.** The algorithm simulates outcomes from the new nodes to estimate their value.
- **Backpropagation.** Results from the simulation are propagated back up the tree, updating the evaluation of earlier decisions.

MCTS could be used to simulate complex decision trees or scenarios, allowing models to learn from synthetic yet structured data. There are, however, some criticisms of this technique:

- **Simplified assumptions.** MCTS relies on simplified assumptions and abstractions to construct its decision trees. These assumptions do not capture the complexity

of real-world systems in the same way as authentic, human-generated data does, and may lead to training sets that lack nuance or fail to reflect true variability.

- **Dependence on input quality.** MCTS heavily depends on the quality of its input parameters and initial assumptions. If simplified inputs contain gaps, the resulting decision trees and simulated data may be biased or unrepresentative, reducing the reliability of the training data. As with all models, "garbage in, garbage out" applies. If the initial parameters or assumptions are flawed, the algorithm may prioritize suboptimal paths, leading to biased or inaccurate outputs.

- **False confidence.** The probabilistic outputs of MCTS can create a false sense of accuracy. The appearance of precision in the results can mask the underlying gaps, uncertainties and flaws in the assumptions and inputs.

- **Computational intensity.** Despite being more efficient than brute-force methods, MCTS requires significant computational resources, particularly for deep or complex decision trees.

What legal professionals should take away from this is that the discussion of these concepts signals a shift in how AI models will be trained in the future, and how liability questions may play out. If pretraining doesn't rely solely on human-generated data, then IP considerations will be due for yet another legal shift.

Balancing Progress with Protection: Takeaways

As we embrace the benefits of AI, we must also grapple with its challenges, particularly around training data provenance, intellectual property and ethical standards.

The way AI models are pretrained isn't so different from how humans acquire knowledge: they process patterns, identify relationships and build from experience. People grow and benefit from their experience––we cannot keep them from relying on it when they make art or generate intellectual property. But there are also temporal and emotional aspects that come along with human learning—we breathe, feel, bleed and suffer as we acquire knowledge and experience. AI, on the other hand, operates at immense scale and speed, eating vast datasets without pause. This ability magnifies its potential, but also its risks. Creating legal and ethical frameworks that govern our use of this technology demands immediate and careful consideration.

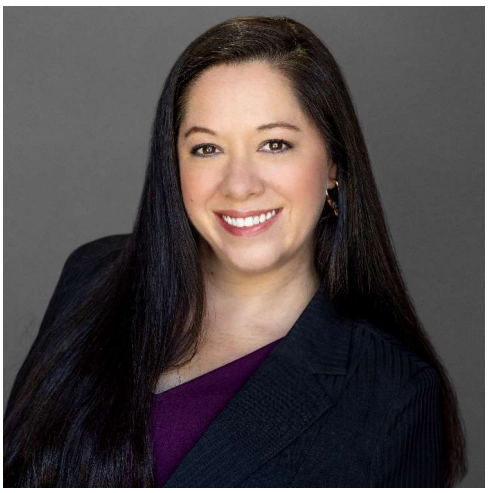
Leading the Charge

As attorneys, we can influence sustainable AI adoption frameworks.

- **Counseling clients.** Emphasize the importance of evaluating the provenance and licensing of training data used to build AI tools, ensuring clients are aware of the associated risks and potential liabilities.
- **Auditing internal tools.** Scrutinize the AI tools used within your own practice to confirm compliance with intellectual property, privacy and data protection standards.
- **Raising awareness.** Dig into how technology works. Share your opinions and insights with others and foster broader understanding about AI's implications across industries.

As regulators demand greater transparency and companies turn to licensed datasets, the legal profession will be essential in shaping how these tools are adopted and used responsibly. Attorneys have a unique opportunity to aim for innovation that balances progress with protection—advancing technology responsibly while safeguarding intellectual property, data privacy and ethical standards. AI may be rewriting the rules, but the legal profession must stay the author of accountability.

Author



Melissa Heidrick

Melissa is the founder of mmData, an organization dedicated to spatial computing for the legal sector. She is also a technology columnist for the American Bar Association's Law Practice Magazine, and serves as Of Counsel for...