Podcast Transcript

EPISODE #272

The Bill is Due: AI Training and Intellectual Property

Published January 23, 2026 • Runtime: 26:19

https://myweirdprompts.com/episode/ai-copyright-data-remediation/

EPISODE SYNOPSIS

In this episode, Herman Poppleberry and Corn dive deep into the "accountability phase" of artificial intelligence, exploring the legal and technical fallout of models trained on "pillaged" data. As we move into 2026, the era of consequence-free web scraping has ended, replaced by high-stakes lawsuits and a frantic search for remediation. The duo discusses the massive shift in the publishing industry, where AI training clauses are becoming as standard as movie rights, and the technical hurdles of "machine unlearning"—the near-impossible task of removing specific data from a pre-trained model. From the "data poisoning" tactics of Nightshade to the architectural promise of the SISA framework, Herman and Corn break down how creators are fighting to protect their intellectual property. They also examine the rise of licensed datasets and the potential for a collective licensing model similar to the music industry. Whether you're an author concerned about your digital twin or a developer navigating the new Data Provenance Initiative, this episode offers a comprehensive look at the front lines of the AI copyright war.

DANIEL'S PROMPT

Daniel

Following up on our recent discussion about pen name publishing, I have a question regarding intellectual property and training AI models. Many publishing houses are now taking a strict stance on AI usage, but for years, major models have been trained on vast amounts of unauthorized data. Given the potential legal liability, what best practices are emerging for the ingestion of IP-protected work into AI engines? Additionally, what remediation is possible for models already in production that were trained on proprietary data without the authors' consent?

TRANSCRIPT

Corn

You know Herman, I was thinking about our conversation from a few weeks back regarding pen names and the whole world of indie publishing. It is fascinating how much that landscape is shifting. But today, our housemate Daniel sent over something that takes that entire discussion into a much more legal, and frankly, more technical territory. He is looking at the intersection of intellectual property and the massive datasets used to train these large language models.

Herman

Herman Poppleberry here, and oh man, Daniel really hit on a nerve with this one. It is the topic that keeps every copyright lawyer and machine learning researcher up at night. We have moved past the initial shock of A-I and into what I would call the accountability phase. For years, it was basically the Wild West. You scrape everything, you train the model, and you ask for forgiveness later. But now, in early twenty-twenty-six, the bill is coming due.

Corn

Right, and it is not just about future training. Daniel specifically asked about remediation for models that are already in production. Models that were trained on what he called pillaged work. It is one thing to say we will be better from now on, but what do you do with a model that already has millions of copyrighted books baked into its weights?

Herman

Exactly. And the publishing houses are not playing around anymore. As Daniel mentioned, we are seeing these incredibly strict contracts where authors have to explicitly sign away or retain rights specifically regarding A-I training. It is becoming a standard clause, much like movie rights or foreign translation rights used to be. But before we get into the fix, we should probably talk about how we got here. Most people hear the term common crawl and think of it as just a library, but it is much more chaotic than that.

Corn

It really is. Common Crawl is this massive, non-profit repository of web crawl data that has been around since two thousand eight. It is petabytes of data. For a long time, it was seen as this great resource for academic research. But when the A-I boom hit, it became the primary buffet for every major lab. The problem is that Common Crawl does not distinguish between a public forum post and a pirated copy of a best-selling novel that someone uploaded to a random server.

Herman

And that is the crux of the issue. These models were built on the assumption of fair use. The argument was that the A-I is not copying the text, it is learning the patterns of language. It is transformative, not derivative. But the courts are starting to look at that differently, especially after the landmark rulings we saw in late twenty-twenty-five regarding the concept of non-expressive use. When the A-I can essentially regurgitate large chunks of a specific author's style or even specific passages if prompted correctly, the transformative argument starts to crumble.

Corn

I remember we touched on this a bit in episode one hundred and five when we talked about A-I benchmarks. If the model has already seen the test data, the benchmark is useless. It is the same thing here. If the model has already ingested the entire back catalog of a major publisher, it is not just learning language, it is competing with the very people who provided the data. So, what are the best practices emerging right now for ingestion? How are companies trying to do this the right way in twenty-twenty-six?

Herman

The biggest shift is the move toward licensed datasets and the implementation of the Data Provenance Initiative standards. You are seeing deals like the ones OpenAI and Apple have been making with major news organizations and stock photo sites. They are essentially saying, we will pay you tens of millions of dollars for the right to use your archive for training. This creates a clean chain of title. If a company uses a dataset that is certified as opt-in only, they are shielded from a lot of that legal liability.

Corn

There is actually a group called Fairly Trained, run by Ed Newton-Rex, that started a certification program for this. They give a seal of approval to models that can prove they did not use copyrighted work without a license. It is a bit like organic certification for A-I. It tells the enterprise customer, hey, if you use our model, you are not going to get sued for copyright infringement down the line.

Herman

That is a great analogy. But for the individual author, like the ones Daniel is working with, the best practice is often about technical signaling. We have seen a massive update to the way robots dot T-X-T works. It is no longer just about whether a search engine can index your site. There are now specific tags for A-I bots. You can tell the G-P-T-bot or the Claude-bot specifically to stay away, even if you want Google to keep showing your site in search results. We are also seeing the adoption of the C-two-P-A standard, which embeds metadata directly into files to prove their origin and specify usage rights.

Corn

But that only works for the future, right? It does not help the author whose book was scraped in twenty-twenty-two and is now part of a model that is being used by millions of people. This brings us to the second part of Daniel's question, which I think is the most technically challenging. Remediation. Herman, is it even possible to untrain a model?

Herman

This is where we get into the concept of machine unlearning. And I have to be honest Corn, it is incredibly difficult. Think of a large language model like a giant vat of soup. You have added salt, pepper, carrots, and onions. Once that soup is cooked, you cannot just reach in and pull out the salt. The flavor of that salt is baked into every drop of the broth.

Corn

So, if my copyrighted novel is the salt, it has influenced the weights of the entire neural network. You cannot just delete a file and have it gone.

Herman

Exactly. In a traditional database, you just delete the record. But in a neural network, the information is distributed. There is no single place where your book lives. Instead, your book helped tweak billions of tiny parameters. To truly remove it, you traditionally had to retrain the entire model from scratch without that data. And when you are talking about models that cost a hundred million dollars or more to train, that is a non-starter for most companies.

Corn

But there has to be some middle ground, right? I have been reading about things like S-I-S-A, which stands for Sharded, Isolated, Sliced, and Aggregated training. It is a way of training models in smaller pieces so that if you need to remove some data, you only have to retrain a small portion of the model.

Herman

Yes, S-I-S-A is one of the leading frameworks for this. The idea is that you divide your training data into shards. You train a separate constituent model on each shard. Then you aggregate them. If a user asks to have their data removed, you only have to retrain the specific shard that contained their data. It is much more efficient, but it still requires a lot of foresight and architectural planning from the beginning. It does not help with the monolithic models we already have in production.

Corn

What about fine-tuning as a form of negation? Could you essentially train the model on a new dataset that tells it, whenever someone asks about this specific book, do not answer? Or better yet, change the weights in a way that counteracts the original data?

Herman

That is often called negative fine-tuning or unlearning through reinforcement learning from human feedback. You essentially penalize the model whenever it starts to output copyrighted material. It is like putting a muzzle on the dog. The dog still knows how to bite, but it is being trained very hard not to. The problem is that researchers have shown these muzzles can often be bypassed with clever prompting or jailbreaking. It is not true unlearning, it is just a layer of censorship on top of the model.

Corn

That feels like a temporary fix at best. It does not actually solve the legal liability of the data being in the weights. If a court rules that having the data in the weights is itself an infringement, then a muzzle is not going to satisfy the law.

Herman

You are hitting on the legal reality there. In twenty-twenty-five, we saw several cases where the discovery process was focused entirely on the training logs. The plaintiffs wanted to see exactly what was in the training set. If the logs show the copyrighted work was there, the defense of we told the A-I not to talk about it might not hold up. This is why we are seeing a rise in what is called vector database filtering.

Corn

Explain that for us. How does that work in practice?

Herman

Well, instead of trying to change the model itself, you put a filter at the entrance and the exit. When a prompt comes in, it is compared against a database of copyrighted signatures. If the prompt is asking for something that is clearly trying to extract copyrighted text, the system blocks it. Similarly, when the A-I generates a response, that response is checked against a massive index of copyrighted material. If the overlap is too high, say more than a few consecutive words or a specific percentage of a page, the output is blocked or rewritten.

Corn

That sounds like the Content I-D system that YouTube uses for music. It is effective for preventing blatant piracy, but it does not really address the style or the intellectual influence that Daniel was talking about. If the A-I writes a story in the exact style of a living author, using their unique world-building and character archetypes, a Content I-D system might not catch it because the specific words are different.

Herman

And that is the billion-dollar question. Can you copyright a style? Historically, the answer has been no. But we are seeing new legislation being proposed, like the No A-I Fraud Act and various state-level Right of Publicity laws, that aim to create a right of publicity for an author's digital twin. This would mean that even if you do not use their exact words, if you are using a model that was specifically trained to mimic them, you owe them compensation.

Corn

It is interesting to see how this is affecting the publishing industry's behavior right now. Daniel mentioned authors having to sign waivers. I have heard that some publishers are even using tools like Nightshade or Glaze on their digital manuscripts before they are even sent out for review.

Herman

Oh, Nightshade is fascinating. For our listeners who might not know, Nightshade is a tool developed at the University of Chicago. It is essentially data poisoning. It makes tiny, invisible changes to the pixels in an image or the characters in a text. To a human, it looks normal. But to an A-I, it is confusing. It might make the A-I think a picture of a cat is actually a picture of a toaster. If enough poisoned data gets into a training set, it can actually break the model's ability to function. It is a way for creators to fight back against unauthorized scraping.

Corn

It is a digital scorched earth policy. If you are going to take my data without permission, I am going to make sure that data hurts your model. It is a very aggressive stance, but you can see why authors feel it is necessary. They feel like they are being forced to provide the raw materials for their own replacement.

Herman

And that brings us to the idea of the collective licensing model. This is something that has worked in the music industry for decades with organizations like A-S-C-A-P and B-M-I. The idea is that A-I companies would pay into a central fund, and that fund would be distributed to authors based on how much their work was used or how much influence it had on the model's outputs.

Corn

That seems much more sustainable than everyone suing everyone else for the next decade. But how do you measure influence? How do you prove that my book contributed three cents worth of value to a specific A-I-generated response?

Herman

That is where the technical side of attribution comes in. There is a lot of research right now into influence functions and Shapley values. These are mathematical ways to trace a model's output back to specific training examples. It is not perfect, but it can give you a statistical probability. It might say, there is an eighty percent chance that this paragraph was influenced by the works of this specific author.

Corn

It is like a D-N-A test for ideas.

Herman

Exactly! And in twenty-twenty-six, we are starting to see the first pilot programs for this kind of automated royalty system. Some of the smaller, niche A-I companies are using it as a selling point. They can say to their users, you can use our A-I guilt-free because we are tracking every generation and paying the original creators.

Corn

I wonder if that will eventually become a requirement for getting insurance. If you are a big corporation and you want to use A-I for your marketing or your internal reports, your insurance company might say, we will only cover you if you use a model that has a verified attribution and payment system in place.

Herman

I think you are spot on. Risk management is going to drive this more than anything else. The fear of a massive class-action lawsuit is a powerful motivator for ethical behavior. But let's go back to Daniel's point about the publishing houses. They are in a tough spot. They want to protect their authors, but they also do not want to be left behind. Some publishers are actually looking into building their own proprietary models trained only on their own back catalogs.

Corn

That makes a lot of sense. If you are a publisher like Penguin Random House, you have a massive, high-quality dataset. Why give that away to a third party when you could create a Penguin A-I that is specifically designed to help your authors with research, editing, or even marketing, all while keeping the I-P within the family?

Herman

It turns the threat into an asset. But it also creates a fragmented ecosystem. Instead of one giant A-I that knows everything, we might end up with dozens of specialized models. You have the Legal A-I, the Medical A-I, the Sci-Fi A-I, each trained on specific, licensed data.

Corn

Honestly, that sounds like a better world. I would much rather use an A-I that I know was trained on high-quality, verified information than one that just inhaled the entire internet, including all the garbage and misinformation.

Herman

I agree. It brings a level of intentionality back to the process. But we have to address the elephant in the room, which is the existing models. What happens to the G-P-T-fours and the Claudes of the world if the courts decide they were built on illegal foundations?

Corn

That is the nuclear option. A court could order a model to be deleted. We have seen this happen before with the Federal Trade Commission in the United States. They have ordered companies like Everalbum and Weight Watchers to delete not just the illegally collected data, but also the algorithms that were built using that data. It is called algorithmic disgorgement.

Herman

Algorithmic disgorgement. That is a heavy term. It basically means you have to vomit up the fruits of your labor if the ingredients were stolen. If that were applied to the major L-L-Ms, it would set the industry back years. It would be a total reset.

Corn

Do you think it will actually come to that?

Herman

It is unlikely for the biggest players. They have too much lobbying power and the economic impact would be too great. What is more likely is a massive, industry-wide settlement. A one-time payment of billions of dollars into a creator fund, combined with a commitment to move toward a licensed, opt-in model for all future training. It is the classic corporate solution. Pay a fine that is large enough to look good in the headlines but small enough not to break the company.

Corn

It is a cynical view, but probably the most realistic one. But for the independent author or the small publisher Daniel is talking about, that settlement might not feel like justice. If you are an author who spent five years writing a masterpiece, and it was used to train a machine that now sells for twenty dollars a month, a small check from a settlement fund might feel like an insult.

Herman

And that is why the best practices for creators are shifting toward defensive technology. Authors are starting to realize that once their work is online, it is vulnerable. We are seeing a move back toward gated communities. More authors are using platforms like Substack or private Discord servers where they can control access to their work. They are moving away from the open web because the open web has become a harvesting ground.

Corn

It is the end of the era of the free and open internet, in a way. If everything you put online is going to be used to train your competitor, you stop putting things online for free. It is a sad side effect of this technology.

Herman

It really is. We are seeing the enclosure of the digital commons. But on the flip side, it is making high-quality, human-curated content more valuable than ever. In a world flooded with A-I-generated text, a book that is verified as one hundred percent human-written and not used for training becomes a luxury good.

Corn

That is an interesting pivot. Maybe the pen name publishing world Daniel mentioned will start using a new kind of watermark. Not just a digital one, but a legal and ethical one. This book was written by a human, for humans, and no machines were harmed or helped in the process.

Herman

I love that. Like the fair trade coffee of literature. But let's get practical for a second. If someone is listening to this and they are a creator, what should they actually do?

Corn

First, update your robots dot T-X-T file. It is the easiest first step. Even if it is not a perfect shield, it sets a legal precedent that you did not give permission. Second, look into tools like Glaze or Nightshade if you are sharing your work digitally. Third, read your contracts very carefully. If a publisher or a platform asks for the right to use your work for machine learning or model training, that is a separate right that should be negotiated and paid for.

Herman

And I would add, keep an eye on the emerging standards like C-two-P-A. This is a technical standard for content provenance. It allows you to attach a permanent, tamper-proof record to your digital files that says who created it and whether it was modified by A-I. As these standards become more common, it will be much easier to prove when your work has been misused.

Corn

Herman, you mentioned earlier that some companies are trying to do this the right way. Are there any specific examples of models that are being trained ethically right now?

Herman

Yeah, there are a few. I mentioned the Fairly Trained certification. There is also a project called Common Voice by Mozilla, which is focused on speech data. They are very transparent about where their data comes from and they only use what people have explicitly donated. In the L-L-M space, there are companies like Big Science that created the Bloom model. They were very careful about the ethics of their data collection, though even they struggled with the sheer scale of it.

Corn

It seems like the smaller the model, the easier it is to be ethical. When you are trying to build a trillion-parameter model that knows everything about everything, you almost have to scrape the whole internet, and that is where the ethics break down.

Herman

Exactly. Scale is the enemy of ethics in this case. But we are also seeing a trend toward smaller, more efficient models that do not need the entire internet. If you are building a model specifically for medical research, you only need medical papers. And you can license those. You do not need to scrape Reddit or a fan-fiction site to understand how a specific protein folds.

Corn

That is a great point. The future of A-I might be a lot of small, clean models instead of one giant, dirty one. I think that would solve a lot of the I-P issues Daniel is worried about. It makes the data ingestion process manageable and auditable.

Herman

Auditable is the key word. In twenty-twenty-six, we are seeing the rise of A-I auditors. These are third-party firms that come in and inspect a company's training data, their model weights, and their output filters. They provide a report to investors and regulators saying, yes, this company is following best practices. It is becoming a standard part of corporate due diligence.

Corn

It is amazing how quickly this whole ecosystem of accountability is springing up. It feels like just yesterday we were all just playing with chatbots and not thinking about where the words were coming from. Now, it is a matter of international law and multi-billion-dollar industries.

Herman

It is the natural evolution of any powerful technology. First comes the wonder, then the abuse, then the regulation. We are firmly in the regulation phase now. And honestly, I think it is going to make the technology better in the long run. It is forcing us to be more precise and more respectful of human creativity.

Corn

I hope you are right, Herman. It would be a shame if the legacy of A-I was just a mountain of lawsuits and the death of the open web. But if it leads to a more sustainable way for creators to be compensated for their work, then maybe it is worth the growing pains.

Herman

Well said. And hey, if you are listening to this and you have thoughts on the matter, we would love to hear from you. We have been doing this for two hundred and sixty-eight episodes now, and the feedback from our listeners is always the best part.

Corn

Absolutely. And if you have a second, leaving a review on Spotify or your favorite podcast app really does help other people find the show. We are just two brothers and a housemate trying to make sense of the world, and every bit of support counts.

Herman

It really does. Daniel, thanks for the prompt. It was a deep one, but I think we covered some good ground. There is no easy answer to the remediation question, but the fact that we are even talking about machine unlearning is a sign of how far we have come.

Corn

Definitely. We will have to keep an eye on those court cases. I have a feeling we will be revisiting this topic before the year is out.

Herman

No doubt about it. Alright, I think that is a wrap for today.

Corn

Thanks for listening to My Weird Prompts. You can find all our past episodes and a contact form at my-weird-prompts-dot-com.

Herman

Until next time, stay curious and keep those prompts coming!

Corn

Bye everyone.

Herman

See ya!