

# The NYT vs OpenAI Is Not Just a Legal Battle

There are three layers to this complex conflict and we must understand all three

[Alberto Romero](#)

January 4, 2023

<https://albertoromgar.medium.com/the-nyt-vs-openai-is-not-just-a-legal-battle-35269cb7065d>



*This article is a selection from [The Algorithmic Bridge](#), an educational project aimed at bridging the gap between AI and people.*

It's 2024 already. Amid New Year resolutions and a newfound determination to defy bad habits and build better ones, the difficulty of leaving behind grudges of the past stands out. It's not that striking, however, when the dispute that's carried over is not just a past grudge but a life-or-death battle that began a week ago.

I'm talking about the [New York Times \(NYT\) lawsuit against OpenAI](#).

Before you log off thinking you already know all there is to know about the topic, let me tell you that this article is not about the legal conflict or even the technical discussions that underlie the NYT's arguments (although I will touch on that).

It's primarily about a deeper fundamental disagreement that separates — perhaps hopelessly — those who support one side from those who support the other: The relationship between morality and progress.

I don't know where I stand yet. Let me be clear on this — I'm not a by-default defendant of the NYT's position just because I'm aware of the practices OpenAI engages in (outlined in the lawsuit).

The topic requires a much closer look than most people are willing to give it. That's what this article is: me thinking out loud about what this means for AI, us, and our future.

### **The legal layer: Not only NYT. Not only OpenAI**

It's also a broader glance at a broader problem.

The NYT is not the only victim affected by the problems it has thoroughly illustrated in the document and neither OpenAI nor Microsoft are the only ones accused of engaging in the practices that moved the NYT to sue them.

The 2023 generative AI landscape has been defined by a large amount of new R&D and deployment initiatives (both open and closed). The only other trend comparable in relevance and intensity has been the number of people denouncing those very initiatives in similar terms as the NYT has done with OpenAI.

Language models are also not the only type of AI system affected. Recent (and not so recent) investigations have revealed that image generators like [Midjourney](#) and video generators like [Pika Labs](#) might have crossed a line as well (there are many more examples like those).

It's not so much the novelty of the dubiously legal practices that have put AI companies in the spotlight but the rapid quality improvements of the tech, which have made the misdeeds more evident. The legal war on generative AI has been cooking for a while — in the past month it reached critical mass.

The NYT lawsuit will be, above all else, the straw that broke the camel's back.

Although the problems are generalized, I will focus on the NYT vs OpenAI (and Microsoft) case. First, because it's the most prominent example — everyone's talking about it — and second, because it's the most relevant one: [As I wrote in September](#), if anyone can take down OpenAI with a lawsuit that's the NYT.

I don't plan to analyze the legal layer (e.g. whether near-verbatim outputs of language models incur copyright infringement, whether the Napster case is a

comparable precedent, etc.). Others have commented, much better than I ever could, on the validity and strength of the claims and the potential outcomes and their effects. If you're interested, I suggest you read [Cecilia Ziniti](#) (and of course [the lawsuit](#)).

For what it's worth, I've skimmed the document and the examples the NYT provides are compelling to my untrained eye — I was surprised they managed to prompt a GPT model to output near-verbatim copies of copyrighted text data. To me, that feels like the most powerful argument the NYT has.

GPT-4 shouldn't do that — not legally (I can't comment on that) but technically. Has ChatGPT/GPT-4 stored copyrighted data somewhere so that it's readily accessible without consent or attribution?

That's part of what I want to focus on instead: the non-legal counter-criticisms from AI people to the criticisms against AI from non-AI people.

### **The technical layer: How does AI work?**

Including the legal layer, which is the most superficial one (not as in “unimportant” but as in “visible”) there are two more layers at which the conversation is unfolding that we should understand as well. Both are used by pro-AI people with variable degrees of success.

The second layer is the technical one, i.e., how GPTs work under the hood. This level of debate is interesting because, on the one hand, lawyers and journalists are often not that well-versed in the intricate structural or functional details of language models, which gives way to [technically trivial counterpoints](#).

And, on the other, although pro-AI people use this knowledge gap to their advantage, most of them don't understand the technology as well as they should to back up their apparent confidence. The reason is simple: Generative AI systems are largely inscrutable, quite novel, and, at times, [eerily alien](#).

This isn't to say we know nothing about what language models are. No one knows exactly how they process the inputs internally to transform them into outputs but we have good conceptualizations of what it's possible or impossible for them to do.

I particularly like [this description by François Chollet](#). Paraphrasing: Being lossy compressors, language models can't possibly store all the data they're trained on. But nothing prevents them from storing a tiny subset of the data in a way that the right input generates an output that's a sufficiently faithful depiction of that data.

In those cases, as the NYT has shown, the output would be, by all accounts, a copy. I.e., an instance of plagiarism and potentially copyright infringement.

ChatGPT might indeed be “a blurry JPEG of the web.” Unsurprisingly, not one equally blurry for all corners of the web. For some portions of the training dataset, there might be no blurriness at all. If that’s the case, then there would be no need to use web browsing or some other kind of retrieval-augmented generation (RAG) mechanism to obtain near-verbatim copies of existing text, as some have suggested.

That’s a potent technical detail that supports the NYT’s legal arguments. (A different issue is whether the user or the company should be held liable, which is not the focus of this essay.)

Interestingly, I don’t think AI evangelists believe these cases are *not* instances of copyright infringement (e.g. because the system is merely taking inspiration from NYT articles, which would be both a misconception of how AI works and a misunderstanding of copyright law).

What they are truly arguing for, if my interpretation of their intentions is correct, is that the law should change to accommodate these cases — and any other potential illegalities AI systems, the companies, or the users might incur, for that matter — whatever the courts ruling happens to be.

I think the anchor that keeps them strongly attached to their views; the reason that forces them to hold onto a stance that appears weak against the existing evidence — both legal and technical — is not that they believe themselves to be legally right or technically more knowledgeable, but something else. Something more powerful.

That leads us to the third and deepest layer of this issue. The one I think requires more press attention and is attractive enough to deserve our sincere consideration.

### **The final layer: The morality of progress**

This layer is concerned with the disagreements about the complex relationship between morality and progress.

I believe the legal and technical arguments expressed by the pro-AI side are merely meant to avoid entering into this underlying complexity that’s harder to defend, which they use only as a last resort.

That's perhaps the most important implication of the multilayered nature of the NYT vs OpenAI battle — that we can safely conclude that people on the pro-AI side of the question don't really think plagiarism and copyright infringement are okay if analyzed in isolation.

Instead, to them what OpenAI and the other AI companies are doing is only acceptable because it so happens to be the collective value that, sadly, we have to sacrifice in exchange for advancing our civilization forward.

It's the lesser of two evils.

If they don't lay out their view like I just did, that's because it's not socially clever to do it. Fighting in legal terms (e.g. that's fair use) or technical terms (e.g. that's not copying but inspiration) is easier. But this is the truest point of conflict — differences about the morality of progress. The law is merely a tool to reflect in formal terms what we consider to be moral for humans and humanity. If progress requires it, they'd say, law must change accordingly.

At the moral level, the conversation unfolds more nuancedly yet much more clearly. First, those who side with the NYT would say:

“What OpenAI does is immoral — beyond whether it's illegal or not — because it's damaging the livelihoods of many creators who are, in a darkly ironic way, the sole reason the technology exists at all.”

Then, those who side with OpenAI respond:

“Actually, the NYT is the immoral one because in selfishly fighting to defend its own IP it's hindering the advancement of an innovation that can greatly benefit humanity. History shows that those who side against the development of technology tend to be wrong.”

And that's the crux, right? It's not a *simple* conflict at all.

A historical lens supports the latter view: progress, especially that which happens thanks to technology, has been the [primary source of well-being improvements](#). And it's not even close. It's a bad idea to oppose its course.

But a contemporaneous lens — looking at the present from itself — yields a different perception: this is all a rather sad spectacle where the train of progress takes with it our species while leaving behind as many individuals and collective values as it must, to keep going.

“That's how things are,” say one people.

“But it’s inhumane,” say the other people.

Unfortunately, both are correct.

### **Making together a better transition to the future**

So that’s the core argument of the people defending AI on the copyright issue and likely in all other past and future issues as well: Even if the NYT is right legally and technically, *it doesn’t matter*.

We should not put barriers to the progress that AI companies are making on behalf of all of humanity.

But we can do a more refined analysis. I said progress through technology yields improved well-being and that’s true. But then other questions arise: Does all technology improve well-being? Does all technology improve well-being equally? Is all innovation, just because it is an innovation, above criticism?

Those who answer affirmatively are people like Marc Andreessen, who eloquently defended this perspective in his [techno-optimist manifesto](#): nothing should stop innovation.

Those who say “it depends,” [are with me](#).

Let’s see if we can find a deeper truth.

I believe technology, progress, and science are the engines of civilization and [have been the primary reasons](#) for the drastic increase in well-being (across pretty much any metric you might choose) for the past two hundred years.

Without technology, we’d be *much worse off*.

That’s clear in hindsight but not so much looking forward. Across the centuries, [fear of technology-driven change has been a constant](#). If you could travel back in time, you’d find critics of the tiniest challenges to the status quo. [Novels, bicycles, teddy bears](#)... you’d find backlash for the most inoffensive and most helpful and useful inventions ever. That’s humans — we hate change. It’s also part of our shameful history. We have time and again unwarrantedly opposed progress moved solely by fear.

But just as it’s a mistake to oppose *all technology* by default, it’s also a mistake to embrace it blindly *at all costs*.

There’s a middle ground between those two stances that we should strive to land on. Many innovations are great and unfairly attacked. At the same time, not all innovations have provided a positive value to society (even after being given

more than enough time to prove that they can be beneficial). A highly consensual example without which we'd be better off is nuclear weapons (not the ability to manipulate nuclear energy, though).

It's hard to evaluate which tech advances go where in the beneficial-detrimental spectrum, but it's harder to defend the view that such a spectrum just doesn't exist because all tech belongs to the net beneficial side.

Social media comes to my mind often when I think about the trade-offs of technology. I can't deny its value (I write on Substack, no less) but after reading the [exhaustive work by Jonathan Haidt](#) and others on the effects of social media on young people (especially teenage girls) it's hard to argue that my being able to write on Substack can ever make up for the damage that Instagram and TikTok are doing to the mental health of our young. They're not the same platform but they're the same *idea*.

From that lens, I can only conclude that social media is a net negative. I don't deny its value but consider it insufficient for the harm it creates.

That's what I dislike about the pro-AI arguments alluding to the morality of progress. They put the hypothetical benefits of generative AI, which hasn't yet proved to be that valuable, above any attempts to steer it to make it more positive while we can (not saying the NYT wants that, though). I'm not talking about attempts at killing it (which I think is absurd) but attempts at making us all collectively better off in the ways that make sense now.

Is it better to let OpenAI sweep away whatever institutional protections it can find in its path for the sake of getting to an uncertain future faster rather than force it with the power of law to pay creators their due for the work they have done, which would slow down AI but not stop it?

And you may say: that makes the technology more expensive for users. Yes, and? I would *gladly* pay twice the cost for GPT-4 if that means OpenAI uses that extra money to license *all* the copyrighted data it uses to train its models (not just the part whose owners they're afraid of). I'd rather have some of my money go to precarious creators than to line, to the last cent, OpenAI's pockets (let alone now that it's a closed, capped for-profit startup).

We like to say that AI is coming, that it's inevitable. And it may very well be (I think so), but it matters how we make the transition to that inevitable future. We're not walking a fixed path. As humanity, we're making decisions all the time.

Decisions about which we have control. Decisions that shape the future we're heading to.

We've never been very caring about that — most people either firmly oppose change or fight for it to happen unconditionally. But we can choose to do it better this time.

If anything, that would make us morally better than our predecessors both in protecting the rights of our contemporaries and ensuring the wellbeing of our offspring.

To me, it makes no moral sense to yearn so much for progress in some aspects if in others, just as important, we regress instead.