PROFESSIONAL BACKGROUND

I am a programmer with several decades of experience (mostly back-end web development, some experienced with compiled languages). I have a Masters degree in Information Security, and have taken in-depth classes on Data Mining (which has seeded the ultimate development of AI) and have written a few algorithms to analyze big data and spit out predictive results.

I am also both a traditionally-published and an independently-published author, and I have a working understanding of intellectual property for purposes of ebooks, physical books, and audiobooks.

ANSWERS AND DETAILED EXPLANATIONS

I will start by offering my recommendations in answer to the primary questions asked. After giving these recommendations, I will explain my reasoning for each one in greater depth.

1. The use of copyrighted works to train AI models.
It is my honest recommendation that copyrighted works should not be allowed in AI training at all. I know this is a strong stance, but it is based on both technical understanding and current, practical market conditions, which present some very unique problems.

In the event that the above recommendation is ignored, I would instead recommend that a) a specific right for AI ingestion is reserved for creators, and must explicitly be signed away, and b) any agreement to submit work for AI ingestion MUST be affirmatively opt-in and NOT opt-out. Opt-out agreements have recently been abused, and do not function well with this specific technology, as I will explain further down.

Market Conditions, Monopsonies, and Coercion.
In theory, there should be nothing wrong with using copyrighted works to train AI models, as long as the owner of the copyright has opted in and signed away a specific right for use in AI training. In practice, many creative markets currently involve publishers and distributors with a monopsony (as the Justice Department has already recognized in the matter of the Penguin Random House/Simon & Schuster). It also bears mentioning that any author with access to their own data will tell you that Amazon currently controls up to 90% of their book income. This state of affairs offers unprecedented leverage to publishers and distributors, who already demand as many specific rights as their current, superior bargaining position will allow them. Adding a specifically reserved right for use in AI data sets will almost certainly result in publishers and distributors simply requiring that right in order for an author to be published at all (buried in a clause in an unexpected corner of their contracts, as has happened now with both Findaway and Apple, before authors belatedly discovered it).

For further consideration in the matter of audiobooks: Even if AI data set ingestion is given its own explicitly reserved right, this right should belong to audiobook narrators and not to authors or publishers, since narrators have the greatest vested interest in the possible duplication or exploitation of their own voice and performance. Authors and publishers should not even be entitled to submit a narrator's work for AI ingestion—and yet, current permission models assume that authors and publishers are capable of opting in on behalf of the narrator, simply by submitting the work for normal distribution. Authors have had their permission to ingest audiobooks assumed and then been asked to email in order to opt out (as in Apple's case). Given that narrators almost never handle submission directly (they are contracted workers who interact with authors and publishers, or, in rare cases, offered a royalty split), why should companies like Apple even assume that authors and publishers have the capacity to consent on their narrators' behalf?

Technical Data Extraction.
Lastly, of course, there is the question of whether data can later be extracted from a model at all. I try to remain humble about my own knowledge of my industry, but in my personal experience, I have never seen a model which could have one piece of data later extracted. In the matter of Apple, which likely uses several automated systems, did the company submit uploaded audiobooks immediately into a data set for a model? In which case, how might a rights-holder even opt out afterwards, as Apple claims is possible?

I reiterate that current market conditions make it functionally impossible for creators to maintain their right to avoid letting their works be ingested. As such, until the market can be better equalized, AI models should not be allowed to ingest copyrighted content at all. This MAY change at a later date.

2. The copyrightability of material generated using AI systems.
Under current market conditions, the results of generative AI should not be copyrightable at all, as long as there is a major AI component to the work. Things which may be copyrightable would include spell-checked or grammar-checked work, which uses a major human component as the base of the work and uses AI only to make minor adjustments. Things which should not be copyrightable include AI generated images and entire AI generated paragraphs.

I have two reasons for this recommendation.

Market Conditions.
The first reason, as mentioned, has to do with market conditions. Because the ingestion of data has not yet been reformed, almost every image, audio, or text which has been generated so far has its roots in some element of ignoring creator rights or else deceiving or coercing creators out of said rights.

Social Impacts.
The second reason has to do with the practical impacts of monetised AI-generated content on our society and existing technology. Already, we’ve seen instances of companies mass-generating “news” or editorial content using AI and putting it up on the Internet in order to make money from it. This material is almost entirely *factually* nonsensical, though AI-generated text is good at sounding authoritative. As such, the Internet is already becoming far less useful, with increasing amounts of effort required in order to sift through bogus AI-generated content. Search engine companies—some of the very companies which most enthusiastically champion AI—do not seem to comprehend that they are making their own job more difficult by doing so. Allowing AI generated material to be copyrighted creates a

brinkmanship problem: Content creators in search of a quick buck waste computing resources by deluging the Internet with nonsensical material, while search engine companies are forced to waste resources creating ever more elaborate methods of sifting through and discarding the junk. Given the sheer resources which AI already requires in order to generate a result, this feedback loop is both wildly wasteful and unsustainable.

This problem will only become worse as time goes on, due to the concept of model collapse. Since AI companies are scraping the Internet without need to worry about copyright, they continue to train their models on data which contains an increasing amount of AI-generated material, itself. This eventually leads to complete model failure—a circumstance which could cripple AI companies, but only *after* they have first made the Internet almost entirely unusable. AI companies know that model collapse is likely, which is why they assign low-wage human workers to comb through material and except data which is obviously vulgar, dangerous, or fake… but even the average person is often taken in by AI-generated articles, which sound authoritative even while they spout complete nonsense. A low-wage worker has little chance of cleaning up this false data unless it is wildly obvious.

From an authorial standpoint, AI generated work is more of a headache than it is an assistant. A writer must work far harder in order to accept an AI-generated story and edit it into something coherent than they do in order to write a coherent story from scratch. Brainstorming with AI can be helpful—but AI-generated work as the basis for an entire copyrighted work should be discouraged for the sake of writer sanity, if nothing else. Studios are obviously very keen on the idea of using AI to generate basic scripts and force writers to edit them into better work. I submit, as both a programmer and a writer, that this is solely because studios are currently run by people who don't understand either writing *or* advanced technology. It's possible, of course, that studios *do* understand that AI-generated work creates more trouble than it solves… in which case, by "creating" a script and then handing it over to an "editor" rather than to a writer (who must do even more work editing the useless script than they would do writing something original), they may therefore pay the "editor" a lower rate. Either way, wholesale AI-generated work provides no actual value to society; in fact, it provides a *negative* value.

3. Potential liability for infringing works generated using AI systems.
Liability for generated infringing works should be equivalent to liability for normal infringing works. There is a saying in information security: "There's no new crime under the sun—just existing crimes committed using new technology." This seems to be another instance of that.

4. The treatment of generative AI outputs that imitate the identity or style of human artists.
Given the explanations of negative social impact I've mentioned above, outputs which imitate the identity or style of human artists should be treated identically to regular copyright infringement. There is absolutely no reason why US copyright should *encourage* the use of AI to generate new monetized content at this time.

OTHER RELEVANT COMMENTARY

I would hate for those reading my comments so far to believe that I think AI is without any merit whatsoever. I *do* believe that AI can serve a useful function. When trained with limited, self-contained, human-verified data and programmed to solve a very specific problem within a very specific domain (for instance, in order to discover new medically-useful molecules), AI can be an invaluable tool.

Companies which claim that AI can be turned into an everything-tool, however, are grossly misrepresenting the technology. While it is technically *possible* that AI may someday become so advanced that it can generate an entire manuscript with greater speed and artistry than a human being, for instance, such an AI model would require wildly enormous resources, and would require constant retraining on human handpicked data in order to stay up to date. Any AI which involves sourcing bulk data from unreliable sources (such as the Internet) is doomed to failure.

This requirement for high resources, limited data, and specific domains is why AI does *not* perform well when generating factual, coherent text for more than a few paragraphs at a time, while it *can* perform relatively well at generating images (a much narrower domain)—at least until such time as model collapse comes for the image generators. As such, while the use of AI should be actively encouraged in certain technical domains, it loses much of its potential social value when applied to creative domains, and in fact generates negative value instead. In short, there is little benefit to be had in privileging AI-generated content in creative contexts at all.