The development of generative AI requires the processing of large amounts of copyrighted materials, including published books, STM (scientific, technical, and medical) and other journals, and news magazines and other news items. The properties and affordances of generative AI depend in part on the size of the dataset. Several large AI companies have already processed very large datasets—and they will need access to more. This is not, however, all a big players’ game. The development of smaller scale generative AI systems based on bases provided by AI companies is likely to flourish in specialized areas (law, medicine, construction, engineering etc.). All those AI users, big and small, will need to find and process large amounts of quality data to create outputs. Add to this that a significant number of users will need to be able to access, replicate, and use datasets and outputs across borders, thus making “national” situations only part of the equation they need to solve.

Looking around the world, there is no uniform standard for how copyrighted materials can be used in AI technologies. This provides opportunities for licensing options that can benefit the copyright system overall. In the judicial system, several lawsuits are pending—perhaps more so in the United States than elsewhere at this juncture given that many AI “giants” are based there—to determine the exact scope of the fair use doctrine. The lawsuits will also determine the extent of users’ liability for the removal (which commonly happens when digital files are processed for machine learning) of copyright management Information (CMI), which, in the United States, can lead to the imposition of statutory damages—even if the copyrighted work is not registered (under 17 U.S.C. § 1202).

Several other jurisdictions have adopted specific legislation to allow text and data mining (TDM), including the EU Directive on Copyright in the Digital Single Market, which contains two exceptions: one allowing TDM by certain research institutions and another for other users, including commercial companies, that is subject to an opt out, the avowed purpose of which is to lead to the formation of a licensing market.¹ The draft AI Act—a proposed EU Regulation (this not requiring implementation in the same way that a directive does)—would obligate “providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video (‘generative AI’) and providers who specialise a foundation model into a generative AI system [...] without prejudice to Union or national or Union legislation on copyright, [to] document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law.”² In parallel, other norms (e.g., FAIR data) are being discussed and developed to require transparency in the use of data to train LLMs and other forms of AI. Japan allows TDM because this type of use is considered not to constitute “enjoyment” of copyrighted works, though the exception is subject to both contractual and technological protection measures (TPMs). The exact scope of exception varies in other jurisdictions (Singapore, Switzerland, etc.). None of them offer cross-border solutions for the simple reason that no national law can.

There are three ways in which the current legal uncertainty will be removed or diminished in a way that will allow machine learning (ML) and specifically AI companies working on generative AI technologies to progress with greater clarity about the metes and bounds of copyright law: (a) legislation; (b) court decisions; and (c) a market-based solution.

In a nutshell, national legislative initiatives such as those outlined above provide guidance within the borders of those jurisdictions. Relying on legislative changes around the globe means that rights holders and users must wait and figure out how to comply with a patchwork of rules.

Those who have pinned their hopes on courts must arm themselves with patience, not to mention the legal fees and risks of an uncertain outcome, which is also likely to vary by jurisdiction. Court decisions will take years, and, in the end, they may not even set clear rules. A future U.S. Supreme Court judgment is expected to happen, but who can predict with accuracy what the court will say? It could create new areas of uncertainty. The bottom line is this: It is extremely unlikely that courts will find, even within one jurisdiction, that all types of AI systems are fair uses and thus fully exempted. Even in a scenario in which a court would find that much of what an AI user has done in jurisdiction x is exempt from liability for copyright infringement and CMI removal (in a way that comports what that jurisdiction’s obligations under applicable international treaties, including the WIPO Copyright Treaty), the cross-border use problem would remain and any copy of the underlying works, even partial, on a server in a different jurisdiction could trigger new lawsuits. Just as importantly, relying on courts to put an end to the debate entails risk for all parties concerned: for users, that a significant part of their use is not exempted by fair use or other legal doctrines or violates CMI-related rights; for rights holders, that overbroad exemptions for entire categories—or very large portions of them—are adopted, shrinking the scope of their rights.

It gets worse. Even if all parties were to agree to wait for courts to solve the issue, a negotiated solution would still need to be found going forward. Given that lawyers are likely to disagree about what a future court opinion will mean (for companies not party to the case), negotiations may not be easier after the issuance of those judgements than they would be now.

Against this background, the indemnifications offered by several major AI providers—and specifically the scope limitations—should be evaluated very cautiously by users of AI systems for at least two major reasons. First, typically, those indemnifications apply only to outputs. If the user makes a copy and/or adds their own training materials to a dataset for machine-learning purposes, this is not covered. Yet, this is precisely the issue at the heart of many of the pending lawsuits. Second, the legal text supporting the indemnifications for possibly infringing outputs often contains significant exclusions. For example, OpenAI’s Terms of Service exclude indemnification for outputs which the “Customer or Customer’s End Users knew or should have known the Output was infringing or likely to infringe, (ii) Customer or Customer’s End Users disabled, ignored, or did not use any relevant citation, filtering or safety features or restrictions provided by OpenAI, (iii) Output was modified, transformed, or used in combination with products or services not provided by or on behalf of OpenAI, (iv) Customer or its End Users did not have the right to use the Input or fine-tuning files to generate the allegedly infringing Output, (v) the claim alleges violation of trademark or related rights based on Customer’s or its End Users’ use of Output in trade or commerce, and (vi) the allegedly infringing Output is from content from
a Third Party Offering.” Whether the “should have known” clause imposes a duty on users to check whether a particular output may be infringing is unclear, but the standard is certainly open to various interpretations. Moreover, excluding any material that the user modified is noteworthy, as many users are likely to tweak the machine’s output. Even a format change might be sufficient to exclude the application of the protection. Indemnifications offered by Google and Microsoft also contain important limitations. Google’s indemnity clause excludes customer uses “after receiving notice of an infringement claim,” for example. There is ample support, therefore, for the claim in a recent Forbes article that “if you read the fine print, the protections offered are narrower than what’s suggested by the PR.”

Finally, there is another crucial aspect to consider. While an AI machine may not “care” about the quality of its training data, users should. To ensure that AI can contribute more positively to progress, AI machines need to learn from high quality data, that is, in many cases, content published by professionals, who are more likely to want to protect it and license it as appropriate. Put simply, all users, not just large ones, should have access to high quality training materials. Otherwise, we may end up with a small oligopoly of very large AI providers—many of whom have already trained their AI systems on almost all available materials, sometimes regardless of quality. These large AI providers may have the appetite to remain involved in long and costly litigation with uncertain outcomes, and thereby forego entering into appropriate licenses that would enable lawful use of copyrighted materials and possibly end ongoing disputes. On the other hand, a diversity of AI models, and providers, built on licensed content would encourage new AI systems that could be developed without threat of litigation and, in turn, also result in licensing revenue for authors and publishers. The result would be a robust, healthy marketplace that fosters ongoing innovation and enables entities and rights holders of all sizes to participate.

Against this background, a market-based solution makes perfect sense. For one thing, a license can cover several jurisdictions—indeed the entire world. This is a major reason why it is happening already, as rights holders and users have negotiated deals. The reality is that they will not fully get there, however. Users are likely to need works from many different sources in many different countries. This matters because using only data from certain parts of the world could easily lead to issues of lack of diversity or even discrimination.

Given that (a) users will be both large and small entities increasingly scattered all over the world, and that (b) rights holders, also big and small, will similarly operate in dozens of countries around the world, it seems highly unlikely that bilateral transactions (one rights holder <-> one user) would be able to solve this within an appropriate time frame and with reasonable transaction costs. It is simply not reasonable to expect a user, in many cases, especially a smaller one, to identify every rights holder in every copyrighted work they want to use (even assuming they can determine what is and is not a protected work) and then contact those rights holders one by one. Nor does it make business sense for each rights holder to have an army of licensing agents dealing with potentially thousands of users around the world, not to mention currency and linguistic barriers, for the full scale of potential uses. Put differently, the licensing market as it now stands means that deals can happen between large rights holders.

---

3 https://openai.com/policies/service-terms, sec. 3(b), updated 6 November 2023.
5 Brad Stone, “AI Legal Protections May Not Save You from Getting Sued,” Forbes, 13 November 2023.
holders and large users, with everyone else potentially left hanging. The difficulties of licensing for all those other situations will inevitably lead to concerns about copyright overall.

The sensible, market-based solution out of this quandary is licensing, including direct and voluntary collective licenses that would bring large and small rights holders to the table and allow users of all stripes to obtain an appropriate license to operate. Voluntary collective licensing would allow for fair payments for authors and other rights holders. It would let AI users use quality works and operate without the sword of litigation, delays, and potential large awards of statutory and other damages now hanging over their heads. Even if licenses did not fully solve the issue of past sins, they would remove liability for the future and pave the way for a negotiation over possible past infringements. Voluntary collective licensing would also allow necessary transparency about the copyrighted works used by AI technologies instead of the current situation marred by over a dozen lawsuits that creates incentives for AI users not to disclose the source of the data they use. This seems a much better and saner way forward than the current jungle of norms, partial exemptions, and lawsuits.