Dear Members of the Committee,

The undersigned organizations represent a broad cross-section of stakeholders dedicated to balanced copyright and a free and open internet, including libraries, civil libertarians, online rights advocates, start-ups, consumers, and technology companies of all sizes. We write to alert you to fundamental problems with AB 412 (as amended), a bill that would create a new requirement in California law to discern and disclose information about uses of in-copyright works in developing generative AI models.

The bill would severely disrupt the operation of federal copyright law – specifically, the fair use doctrine, a pillar of balanced copyright law – to an extent that it would be successfully challenged as a violation of the Copyright Act and the Constitution. Indeed, the bill is premised on an untested interpretation of copyright law that is the subject of ongoing litigation: that copyright holders can make a valid infringement claim based solely on AI model training. The bill would thus put draconian regulations in place without a clear, legitimate basis in federal law. Meanwhile, this impossible new regulatory regime would crush the developers poised to make California a global leader in the most important technology of the 21st Century. Requiring AI developers to disclose every copyrighted work their models may have encountered is akin to demanding that every human author—journalists, columnists, novelists—list every article, book, or idea they've ever read that may have shaped their thinking. Such a mandate would undermine both free expression and the creative process itself. We urge you not to move forward with this ill-conceived legislation.

AB 412 is preempted by federal copyright law and would be void upon enactment. The bill's requirements "would frustrate the operation of…federal copyright law by interfering with the exercise of the statutory privilege of fair use," rendering the law void under federal conflict preemption.[1] AB 412 is also expressly preempted by 17 U.S.C. § 301, which establishes that the subject matter and activities regulated by federal copyright law are "governed exclusively by this title" and that "no person is entitled to any such right or equivalent right in any such work under the…statutes of any State." AB 412 would thus violate the U.S. Copyright Act. It would also violate the Constitution's Copyright and Commerce Clauses.

AB 412 interferes impermissibly with federal court consideration of how copyright applies to AI training. Federal courts are working through a wide variety of cases that will further explicate how and why AI training qualifies as fair use under federal copyright law. AB 412 would disrupt the ongoing development of federal copyright law.

[1] *See X Corp. v. Bright Data Ltd.*, 733 F. Supp. 3d 832 (N.D. Cal. 2024).

Compliance with AB 412 is impossible. AB 412 places impossible demands on AI developers, and amendments to limit its scope to works registered with the United States Copyright Office and mandate the use of so-called "approximate fingerprint" technology are fig leaves that do little to mitigate the impossible demands AB 412 makes on AI model developers.

The bill requires model developers to catalog all registered in-copyright works that they "know" were used in the course of training, independently of any copyright holder request. The vast majority of works created in the last century are copyrighted and hundreds of thousands of works are registered every year, so every model developer will "know" that substantial numbers of covered works are in their training data. What kind of knowledge is required to trigger the obligation to catalog a work and its owner? It may not matter, because model developers are also required to "[m]ake reasonable efforts to identify and document any other copyrighted materials [i.e., materials they *don't yet know about*] that were used by the developer to train the GenAI model." What are "reasonable efforts" in the context of this cutting edge technology and the mountains of data used in training? What is reasonable in light of the speculative nature of rights holder legal claims under federal law? These issues would have to be sorted out in the inevitable gold rush of litigation spawned by this bill.

In any event, the bill requires some level of proactive searching and filtering by model developers before any request has been made. That means that the "fingerprint" technology referenced elsewhere in the bill, whatever its utility, is not available to the developer when it makes this initial catalog. The registration records on file with the Copyright Office would be the only authoritative source of information about covered works and their owners, and these records are not machine-accessible or machine-readable.[2] Thus it would be impossible to automate searching these records at the scale required to support analysis of massive AI training datasets.

More importantly, even if registration records were machine-accessible, these records don't provide enough information to enable the identification of relevant works in a training dataset, or the owners of those works. Copyright registration records are akin to cards in a library card catalog and include only basic bibliographical information - a work's title, author, date of first publication and the like. While some registrations are accompanied by deposit copies of works, those copies are not made publicly accessible or searchable online, so a researcher has no way of connecting a registration record with the attributes of the registered work (the words in an article, the appearance of an image, the sound of a recording, etc.). Several types of works commonly found online (like photographs and periodicals) are typically registered in group registrations that can list anywhere between 10 and 750 individual works each, again providing only titles or brief descriptions for the underlying works. Copyright owners are under no obligation to include unique identifiers or even basic bibliographical information with published copies of their works, nor are they required to update these records to reflect changes in ownership. Thus there is no way to reliably match copies found on the open internet to particular works, owners, or registration records. AB 412 makes an impossible regulatory demand of AI

[2] *See* U.S. Copyright Office, Circular 23: The Copyright Card Catalog and the Online Files of the Copyright Office, *available at* https://www.copyright.gov/circs/circ23.pdf.

developers: proactively to ascertain and provide information that simply cannot be obtained at a reasonable cost based on publicly available information.

AB 412 is a solution in search of a problem. AB 412 purports to help copyright holders enforce their rights, but the federal rules of civil procedure already provide anyone who wants to pursue an infringement claim with ample opportunity to prove their works were used by defendants. Copyright holders pressing claims against AI developers have not faced any particular difficulty in proving their works were used, and courts have shown they can adapt discovery procedures to fit new technologies.[3] The composition of major AI training datasets is well-known and most are composed of materials posted publicly on the internet; an owner whose work was published online can be sure it has found its way into many AI training datasets. The presence of a work in a model's training data can also be inferred when model outputs are strikingly similar to a copyright holder's work.

Attempting compliance with AB 412 would stifle competition in AI development and make California radioactive to AI developers. No AI developer could satisfy AB 412 comprehensively, but small developers would be the most impacted by attempting even partial compliance, as they have the fewest resources to spare. Creating barriers to entry for new AI companies stifles competition, reducing innovation and raising costs for consumers. Ultimately, the burden of compliance would lead to companies leaving California or failing to offer their services in the state, hurting California's creative and technology industries and its overall economy.

Transparency policies should serve and be tailored to a clear public interest. Valid concerns about bias, public safety, and other issues may support transparency regulations for technology development. However, such provisions should be carefully tailored to serve a bona fide public interest and to impose reasonable burdens proportional to that interest. AB 412 fails that test; its disclosure requirements are onerous, and its objective is to serve one commercial sector at the expense of another, with no basis in the Constitutional purpose of copyright law, which is "to promote the Progress of Science."

The bill would have a chilling effect beyond large-scale commercial uses. The line between "commercial" and "non-commercial" use in AI development is thin and porous, as many foundational open-source models are developed or improved by commercial entities and then used by non-commercial and research users. Hobbling AI development by commercial entities will thus have immediate harmful effects on non-commercial users, researchers, consumers, and other stakeholders.

---

[3] *See* Jonathan Band, *Presumption of Copying in AI Training*, Disruptive Competition Project (Sept. 23, 2024), https://project-disco.org/intellectual-property/presumption-of-copying-in-ai-training/ (explaining that "courts not only have created reasonable presumptions easing the burden on plaintiffs to prove that copying occurred, they also have demonstrated their adeptness at updating those presumptions to accommodate new technologies.").

AB 412 would unleash a tsunami of copyright trolling and shakedown lawsuits. Every major AI model was trained in part on crawls of the open web, so everyone who has ever posted anything online will have a new cause of action under this bill. Materials posted by individual creators and non-commercial actors online are also the hardest to identify and associate with owners. It's a recipe for trolling and shakedown operations. Registration is hardly a barrier to would-be trolls - for as little as $45,[4] the owner of a single comment posted online could register their work, file a demand with an AI model developer, and start the clock on their $1000/day damages.

Forcing companies to reveal their sources will make it harder for small companies to compete. Identifying and sourcing unique training data can be a source of competitive advantage for firms that otherwise can't compete with the raw power and resources of larger firms. Strategic demands for "transparency" disclosures could be a cover for corporate espionage that reveals small developers' trade secrets, erasing their competitive advantage.

The legislation would create privacy risks. Disclosure demands could lead to the sharing of private information that would not have been disclosed in the ordinary course of AI development and use. As currently drafted, the bill offers no protections against disclosure of this kind, an additional risk to the public interest.

For these reasons, we strongly urge you to oppose AB 412, an unlawful, unworkable, and unnecessary intrusion into the federal copyright system that will harm California creators, consumers, and companies.

Respectfully,

Brandon Butler
Re:Create[5]

Jennie Rose Halperin
Library Futures, NYU School of Law

Meredith Rose
Public Knowledge

Michael Petricone
Consumer Technology Association

Jill Crosby
Engine

Anna Tumadóttir
Creative Commons

Wayne T. Brough
The R Street Institute

[4] *See*, U.S. Copyright Office, *Circular 4: Copyright Office Fees*, https://copyright.gov/circs/circ04.pdf.
[5] Not every member of the Re:Create Coalition necessarily agrees on every issue, but the views we express represent the consensus among the bulk of our membership.