

Commentary

The 1 (Trillion): Why an Obscure Copyright Acronym Could Decide Europe's Future Competitiveness



Caroline De Cock



Remy Chavannes

In the first days of the COVID-19 pandemic, before much clinical literature existed on the new virus, BenevolentAI used a knowledge graph built from machine-read biomedical text to identify existing drugs that might work. It proposed baricitinib, a rheumatoid arthritis drug, because it could block the virus's entry into lung cells and reduce the inflammatory response. [Baricitinib became](#) the first immunomodulatory COVID-19 treatment to receive full U.S. regulatory approval and was later recommended by the World Health Organisation.

No single scientific paper contained the answer. It emerged from relationships across a body of literature, when nobody could have known which papers or publishers would prove relevant. This is precisely what the text and data mining (TDM) exception in Article 4 of the copyright in the digital single market directive permits: extraction by any lawful user, including for commercial purposes, unless a rightholder reserves the work in machine-readable form through an opt-out.

The current policy debate on the TDM exception focuses almost entirely on how it benefits (American) technology companies, at the expense of (European) rightsholders. That framing overlooks how widely TDM is already used in Europe. A legal and factual [analysis](#) prepared by Brinkhof found that European businesses across multiple sectors rely on it for work far removed from the chatbots highlighted in the media. Automated data ingestion supports analytical mining, functional models and generative models. Narrowing the horizontal TDM exception in response to concerns about generative AI would therefore hamper activities that produce no substitutive creative output and that are vital to Europe's industrial competitiveness.

Baricitinib was not an isolated example. In pharmaceuticals, [Bayer](#), Sanofi and [Novo Nordisk](#) analyse millions of biomedical papers from thousands of publishers and open repositories to find drug targets and disease associations. The use of TDM enables to perceive what no researcher can surface by hand, as the useful connection often lies between papers owned by different publishers and breadth of coverage is part of the method, not a convenience. Machine translation works the same way. [DeepL](#), the German service that competes with far larger general-purpose models, is trained on roughly a billion bilingual sentence pairs drawn from multilingual websites, news and other online material, a volume no company could clear source by source. [Heuritech](#) analyses around three million social-media images, against a database of some 500 million posts, to forecast demand for houses such as LVMH, Prada and Moncler and retailers such as H&M, so that less stock is made and less fabric wasted.

And in finance, asset managers including Amundi, DWS and Robeco, alongside analytics vendors such as [SESAMm](#), mine news sentiment, pricing signals and ESG disclosures across billions of public articles both to build strategies and to meet the EU sustainability-reporting duties that other European law now imposes on them.

The public-interest uses are equally broad. Healthcare developers such as Corti, Nabla and Ada Health train clinical assistants and diagnostic tools on medical literature, guidelines and multilingual records, work that matters most for smaller European languages without comprehensive commercial datasets to license. Education companies such as StudySmarter and [Knowunity](#), which serves more than 20 million students, use mining to align tutors with local curricula, while plagiarism services cannot function unless they can index the material they check against. Cybersecurity is perhaps the clearest case: threat-intelligence firms such as CybelAngel and Sekoia.io scan websites, code repositories and illicit forums for leaked credentials and planned attacks. One CybelAngel investigation found more than 45 million medical images exposed on unsecured servers after scanning about 4.3 billion IP addresses and had them reported. None of this can be licensed in advance and all of it helps European organisations meet duties under rules such as the NIS2 Directive.

No Europe-wide licensing mechanism covers the thousands or millions of rightholders involved. Yet these sectors, and European industry more generally, were absent from the European Parliament's March 2026 [resolution](#) on copyright and generative AI, which asked the Commission to assess legal uncertainty and competitive effects before reopening the framework. Its list of stakeholders included researchers, universities, libraries, AI developers, news outlets and the creative sector. The omission of the wider industry exposes a glaring blind spot: copyright is still debated as a niche concern even when changes to its rules could reverberate across entire sectors of the economy. BUSINESSEUROPE made much the same point in its paper [Copyright: A Business Perspective](#): every company in a knowledge economy relies on text, data and software. But they are weirdly absent in a debate that threatens their ability to post-train, domain-adapt and otherwise develop AI in Europe.

Requiring European businesses to clear rights across millions of inputs would create a familiar economic problem: the tragedy of the anticommons. The medieval Rhine provides the classic analogy. Each baron controlling a stretch of riverbank could charge a reasonable toll, while the combined tolls made the river unnavigable. A forthcoming economic analysis for the Lisbon Council by Bartłomiej Biga estimates the scale of the modern risk. Under the current opt-out framework, AI adoption could add about EUR 942 billion to EU gross domestic product by 2035. Requiring prior authorisation for every input would jeopardise much of that value.

The effect of restricting training material is not proportional to the amount removed: excluding a small quantity of high-value material can cause a much larger decline in quality. The Netherlands offers a practical warning. GPT-NL adopted a licensed-only sourcing policy, leaving its Dutch-language training corpus about [90 per cent smaller](#) than the material lawfully available under the existing exception. On factual-knowledge benchmarks, it performed close to random guessing and well below comparable European models trained on broader material. Mandatory licensing would also favour the few companies that might be able to pay for and administer thousands or millions of licences, while the 30-person firms cannot.

Europe's competitors have generally chosen broader access. The U.S. permits commercial mining under fair use; Japan allows use where the purpose is not enjoyment of a work's expression; Singapore prevents contracts from overriding its mining exception; and China supports model development without a comparable copyright restriction. Research teams, computing contracts and investment can all move when legal certainty disappears.

This is the harder truth about competitiveness. It is difficult to name policies that would reliably make Europe more competitive; economists and governments have argued over the recipe for decades. It is far easier to name the ones that would make it less so, and reopening a settled, horizontal exception that already carries drug discovery, cybersecurity, translation and half a dozen other industries is high on that list. Not doing avoidable harm is where a competitiveness agenda begins.

the**Forum**

The Forum is a semi-weekly newsletter of cutting-edge blogs devoted to policy proposals and up-to-date political analysis. It is hosted by the Lisbon Council, a global think tank based in Brussels. For more commentary visit: <https://lisboncouncil.net/category/blog/>