The Coalition for Creativity (C4C) is a broad-based coalition that seeks an informed debate on how copyright can more effectively promote innovation, access, and creativity. C4C brings together libraries, scientific and research institutions, digital rights groups, technology businesses, and educational and cultural heritage institutions that share a common view on copyright. See our members here: https://coalition4creativity.org/about-us/.

General remark 1 - Cautionary tales from other geographies: the case of the Database Directive and the ongoing AI Act tergiversations

A. The Database Directive: a failure that refuses to go away

When considering new copyright legislation applying to data, the European Union 1996 Database Directive[1] is a perfect illustration of a counter productive set of provisions.

The Database Directive has illustrated that once new rights are attributed they are almost impossible to repeal, even if they end-up falling short of their objective.

In 2015, the European Parliament already pointed out that this Directive is "an impediment to the development of a European data-driven economy", and called upon the EC to abolish it (§108).[2]

Dr Annemarie Beunen pointed out that the sui generis right is arguably over-strong and discriminates against both competitors and non-profit users such as researchers, educational institutions, libraries, museums and archives" and adds that it "fails to balance the commercial interests of database producers against the public interests of society at large, such as the free flow of information" (pp. 270-1).[3]

This is usually the case when new rights are granted, namely: they benefit the bigger players who are already active in the market. BEUC, the European consumer organisation, made this point in its response to the EC's first public consultation back in 2006, as it explained that this Directive also impacts consumers because: "it increases the barriers to entry for potential competitors in the database market, thus raising costs and stifling innovation to the detriment of, amongst others, the consumer." In a similar vein, Professor James Boyle, of the Duke University School of Law, has cautioned that "setting intellectual property rights too high can actually stunt innovation."[4]

Dr Alina Trapova, Lecturer in Intellectual Property Law at the University College London Faculty of Laws, has perfectly encapsulated this approach when she stated that: "We have examples in our EU IP [laws] where we have killed a fly with an elephant gun. The Database Directive is one prime example."[5]

[1] http://data.europa.eu/eli/dir/1996/9/oj
[2] https://www.europarl.europa.eu/doceo/document/A-8-2015-0371_EN.html
[3] https://scholarlypublications.universiteitleiden.nl/handle/1887/12038
[4] https://www.ft.com/content/99610a50-7bb2-11da-ab8e-0000779e2340
[5] https://informationlabs.org/podcasts/ai-lab/alinatrapova-15june23/

B. The AI Act: a product safety initiative with inappropriate copyright creep

The European Union is currently adopting an AI Act, originally intended as a horizontal legislation looking at artificial intelligence from a product safety angle.

As is often the case, copyright has emerged in the discussions under the pressure of a variety of rightholders, with a transparency provision being put forward by some members of the European Parliament to the effect that AI foundation models could be required to publish separately information about the copyrighted material they used in their training data, a requirement that both contradicts the existing copyright legal framework and that is impossible to comply with considering the scale of the data included in AI data sets, their origin in certain cases (e.g. the web) and the fact that there is no comprehensive, reliable and up to date information to identify the fact that an item is copyrighted, let alone its underlying rights ownership.

General remark 2 - The competitive advantage of countries that have adopted open norms such as fair use

The open norms approach adopted by the US with "fair use" has shown its continuous capacity at adapting to technological changes and creating an environment that balances the need to encourage creativity whilst not hindering innovation. Over the past decades, various countries have adopted a similar flexible approach to legislating copyright, with substantial benefits to their economic growth and a much more robust legislative framework when it comes to handling a changing environment. This is notably the case in Canada, Israel, Japan, Korea, Singapore, and Sri Lanka.

The European Union works with an exception and limitations to copyright system that is clearly showing its shortcomings in addressing technological change: while the Copyright in the Digital Single Market Directive[6] was adopted in 2019 and is yet not fully implemented in all EU countries, copyright concerns have again been raised in the framework of the AI Act negotiations, even though the adopted copyright legislation comprised articles addressing text and data mining for both commercial and non-commercial purposes. This constant demand for legislative interventions creates an unnecessarily uncertain legal framework that is not conducive to innovation.

Training: copyright protectionist considerations should not be at play at the input level

At its essence, copyright serves as an economic regulation designed to encourage creators to generate and distribute new forms of creative expression. Its fundamental purpose is to benefit the public by advancing the "progress of science," in line with the U.S. Constitution.

As a result, we believe that the evaluation of new technology should primarily focus on its contributions to these objectives, rather than the incidental mechanical or technological methods it employs to attain them.

[6] http://data.europa.eu/eli/dir/2019/790/oj

Looking at the US legislative framework, the use of copyrighted works as training data for generative AI tools should generally be considered fair use, as this principle allows for copying for non-expressive (or non-consumptive) uses.[7] Machine learning is about collecting huge data sets, cleaning them up, chopping them in small parts referred to as tokens, splitting them into training and test data, and allowing them to be extracted in response to a prompt. Depending on the prompt, different tokens can be relevant while others might be completely disregarded. Copyrighted works used as input in AI models become part of a data set. They become tokens that integrate in a data collection and copyright protection does not apply to facts or ideas.

As aptly stated by the Authors Alliance, "If it did, copyright law would run the risk of limiting free expression and inhibiting the progress of knowledge rather than furthering it."[8]

Transparency

Requiring transparency based on untransparent data creates an impossible to comply with threshold.

No one knows what is copyrighted or not. Copyright is not vested upon a work through a deliberate act like a registration: it is bestowed on any creation that meets the requirements of copyright laws, and those requirements may vary from one country to another.

There is no register of copyrighted works and hence there is no way to list separately which of the elements in your data set are copyrighted. For this reason, any transparency obligation that creates a subset requirement for copyrighted works is a recipe for compliance failure through no fault of the entity trying to comply, and hence a recipe for legal uncertainty.

Additional questions that are out of the copyright scope

Questions 30 to 34 put forward in the US Copyright Office notice extend beyond the scope of copyright and should hence not be addressed by the US Copyright Office.

[7] Sag, Matthew, Orphan Works as Grist for the Data Mill (August 30, 2012). Berkeley Technology Law Journal, Forthcoming, http://dx.doi.org/10.2139/ssrn.2038889
[8] https://authorsalliance.substack.com/p/copyright-and-generative-ai-our-views