Copyright Conundrums: Rethinking Authorship Rights Against AI in the Open Access Commons

Leo Lee, Dr Simon Rowberry, Dr Laura Dietz

Introduction

The ‘existential threat’ underlying the rapid development of AI has characterised debates pertaining to the ‘death of the author’, engendering concern for equitable access, distribution of data and those engaging in intellectual work, including the present author.

A fundamental tension resides between ensuring compensation to authors of original works and granting generative AI sufficient corpus to offer its innovative services. Acknowledging that existing regulations primarily focus on rightsholders without addressing impacts on the data ecosystem, the author further highlights the issue of governing open access commons.

Legal Framework of Intellectual Property Regulations

AI is prone to copyright infringement, particularly regarding the legality of web scraping technology in the ‘text and data mining’ (TDM) practices of LLMs which remains dubious.

a) The United States and the ‘Fair Use’ Doctrine

In the United States, using protected material to train algorithms may violate reproduction rights under 17 U.S.C. §106(1), though the 'fair use' doctrine provides an exception. ***Campbell v Acuff-Rose Music*** notes the “transformative” criterion that reduces the relevance of commercial incentives against a rulling in favour of fair use. The threshold is whether the work has **“superseded” the original creation ‘with [a] further purpose or different character’.**

OpenAI claims its training practices are “transformative” on two grounds **1) its practices are “non-expressive” and 2) its output is different.** The first claim suggests that protected works are meant for appreciation of the author's expression. The second posits that the output share limited commonalities with one particular source. However, consider outputs that use protected works with the intent to replicate.

b) The European Union Artificial Intelligence Act 2024

The Act reinforces preexisting copyright law. The first consists of **general copyright protection**, including exclusive rights for temporary or permanent reproduction **(Legal protection of databases Directive).** The second establishes ***sui generis* rights** which gives authors ability to authorize or prohibit reproduction by any means **(InfoSoc Directive).** When a work lacks protection, consider ***Ryanair Ltd v PR Aviation BV*** where owners are entitled to introduce contractual restrictions (terms and conditions), even to prohibit web scrapping when its content is not *per se* protected.

Statutory exceptions to this is noted in **TDM exceptions of the DSMD**. For commercial TDM, Art.4(3) introduces an **‘opt-out provision’** that enables rightsholders to reserve the content of their works for TDM purposes expressly—to negotiate licenses with AI developers or otherwise.

However, Art.7(2) incorporates three step test from the **InfoSoc Directive**: **(1)** certain special cases; **(2)** without conflict with normal exploitation of the work; and **(3)** no unreasonable prejudice to the rightholder. This suggests that the exception cannot legitimize reproductions that compete economically with the protected material used for AI training.

Nonetheless, the provision is narrow considering the multiplicity of rights. Scholars critiques that the provision **neither offers sufficient legal certainty** for users of copyrighted material given the vagueness of ‘normal exploitation’ and ‘legitimate interest’ **nor provides sufficient flexibility** to accommodate emerging cultural and socioeconomic needs for data mining. In ***Pelham v Hütter and Schneider-Esleben***, the CJEU barred the external balancing of competing fundamental rights beyond the three-step test; thereby restricting avenues of discretion.

Framework

Prior scholarship have tended to restrict their scope to copyright law lacking consideration of its associated risks against the greater data commons. This study thus uniquely situates academic discourse towards the sustainability of the open-access data commons.

The research will consist of two constituent sections: (1) Survey major statutory exceptions in the United States and the European Unions; and (2) Regulation of generative AI in relation to its complementary benefits to open-access models and its impact on the data ecosystem.

Generative AI and Open Access Commons

Generative AI poses systemic risks to information integrity due to its low-value, biased and inaccurate content. This engenders a negative feedback loop where poor data becomes future datasets for training LLMs, worsening output as AI perpetuates a self-replicates easily accessible bot-generated content. Additionally, authors are disincentivised from contributing to the digital commons provided that their work is susceptible to exploitation in an ambiguous regulatory landscape.

A potential solution is managing resources through ‘symmetric use privileges’ for all, exemplified by Creative Commons licensing. This regime serves three purposes: **(1)** outlining permitted uses; **(2)** requiring attribution; and **(3)** mandating the continuation of the license for modified works. The regime thus requires 'replenishing' the commons by ensuring quality outputs derived from protected works.

Nevertheless, AI struggles to recognize all works, as its generation relies on billions of existing works. Certain authors **analogised data training to organic learning**. Thus, when output lacks attribution to a unique author, copyright law is unable to enforce open-access licensing obligations.

Szkalej and Senftleben propose a **contractual approach that leverages the ‘opt-out regime’** to require AI developers using CC-licensed works to accept associated obligations. Creators can use infringement as leverage for negotiations, establishing a **‘contractual chain-binding model’** that incentivizes developers to embed CC obligations in their terms. This approach ensures that LLMs adhere to licensing conditions benefiting the rightsholder, regardless of whether the output contains copyright-protected material.

Nonetheless if the output lacks copyrightable elements, CC licensing cannot be enforced, as there is no original work to protect. There is perhaps merit in reimagining consent via **‘social licenses’** for the participatory engagement engender trustworthy industry practices. Developing practice whereby users cite works according to the mandatory guidelines of generative AI platforms, in dialogue with authors, would mitigate users to circumventing licensing objectives in maintaining the integrity of the open access data commons.

Student: Leo Lee
Department: UCL Department of Information Studies, HKU Department of Law
Host unit of research project: Department of Information Studies, University College London, United Kingdom
Supervisor: Simon Rowberry and Laura Dietz (UCL), Haochen Sun (HKU)