Copyright and the Dynamics of Innovation in Artificial Intelligence*

Christian Peukert
University of Lausanne, Faculty of Business and Economics (HEC)
christian.peukert@unil.ch

Abstract

Copyright can determine access to data, which may, in turn, affect innovation in Artificial Intelligence (AI). In this paper, we conduct an empirical analysis to investigate the relationship between copyright and the dynamics of AI innovation. Some countries provide exceptions in copyright law that are relevant for AI, particularly concerning text and data mining and the doctrine of fair use. The study takes a global perspective, comparing different countries based on the breadth of exceptions in copyright law. We find that countries with broader copyright exceptions tend to exhibit higher levels of AI innovation, as measured in a larger number of AI research publications, more active participation in open-source AI projects, increased AI patent filings, and a higher rate of AI venture formation. These findings suggest that statutory provisions enabling data access can be pivotal in fostering AI innovation. However, while broad copyright exceptions may support current AI innovation by allowing the use of existing works, they might also inadvertently discourage the creation of new works, leading to negative long-term consequences for the availability and quality of training data. We discuss how mandatory licensing may provide a setting that aligns the dynamic incentives of rightsholders and AI developers.

Keywords:

AI, Training Data, Copyright, Licensing, Policy

1. Introduction

Data is the key input factor in Machine Learning (ML) and Artificial Intelligence (AI) technologies that enable societal and economic progress, either by discovering knowledge (Bianchini et al., 2022), increasing productivity (Wu et al., 2020), or by expanding existing or serving new markets (Bessen et al., 2022). The perhaps most impressive developments have occurred with the advent of generative AI, opening up new possibilities to produce text, images, and audio. With applications like ChatGPT, Copilot, Gemini and Firefly, AI has become a new tool to express ideas for hundreds of millions of users.

Any ML technology depends on input data and is often trained on information available on the public Internet, including unstructured data in the form of text, images, audio and video. However, because input data may be subject to intellectual property protection, copyright law has emerged as a critical area of the regulatory discourse (Samuelson, 2023). In an interconnected digital world, copyright issues now have global reach, leading a consortium of legal scholars to warn that "outdated copyright laws around the world hinder research" (Fiil-Flynn et al., 2022). For example, in the EU, although copyright was only recently modernized with the Directive on Copyright in the Digital Single Market in 2019, discussions about further reform are in full swing as part of the discussion of new legislation on the regulation of AI. Internationally, we currently witness several lawsuits against developers of AI systems that allege, among other claims, direct copyright infringement by creating unauthorized copies of their works and using such copies as training data. As a result, policymakers worldwide are tasked to balance the goal of innovation in AI and the interests of rightholders. However, a comprehensive review of the academic literature revealed that the available evidence is insufficient to inform and guide policy (Peukert and Windisch, 2024).

In this paper, we explore how differences in copyright law across countries relate to differences in AI innovation across countries. The specific research question we aim to address is: *What is the relationship between statutory access to data and innovation in AI?*

The legal landscape surrounding data usage for AI is complex and varies significantly across jurisdictions, influencing how researchers and developers can access

* I thank Jérémie Haese for research assistance. This research has benefited from the financial support of Google. The study was conducted independently, and the views expressed in this article are mine. Google did not have the right to prepublication approval.

URI: https://hdl.handle.net/10125/109384
978-0-9981331-8-8

and utilize data. Data access can be governed on a contractual basis (via licenses) and on a statutory basis (via exceptions in copyright law). Licenses can range from highly permissive, allowing broad use and modification, to rather restrictive, limiting use to non-commercial purposes or requiring derivative works to be shared under the same terms. Statutory rules typically leave no discretion to individual rightsholders, but lawful access is governed directly by exceptions in the relevant law. So-called Text and Data Mining (TDM) exceptions in copyright law are particularly relevant for AI, as they enable the extraction and analysis of large datasets for research purposes without the need for explicit permission from rights holders. Similarly, some countries enable the lawful re-use of data, for specific purposes and under specific conditions, based on the fair use doctrine. Importantly, jurisdictions vary widely. Copyright law in some countries offers broad exceptions, whereas other countries do not allow the usage of copyrighted material for AI without explicit permission.

To provide an answer to our research question, we examine correlations between statutory access to training data enabled by exceptions in national copyright law and several measures of AI innovation. We compare jurisdictions around the globe with varying degrees of access for non-commercial and commercial research.

We find that countries with broader exceptions exhibit higher levels of innovation activity, with faster growth rates in AI-related publications (on arXiv), open-source software contributions (on GitHub), patents (filed with the USPTO), and ventures (listed on Crunchbase).

Our work builds on existing studies on the economic trade-offs between copyright protection and data-driven R&D. Prior work has shown a correlation between TDM exceptions in copyright law and related academic research (Filippov and Hofheinz, 2016; Palmedo, 2019; Handke et al., 2021). Regions that require specific rights holder consent for TDM see a lower share of TDM-related research in overall publications, and even more so in regions with strong rule of law. However, prior evidence is limited to only counting the number of publications that feature the word "data mining" in the text. It also only captures publications that predate important breakthroughs in ML and AI.

By providing empirical evidence on the relationship between copyright law and AI innovation, our study offers timely insights for policy. Overall, our results suggest that statutory access to training data can be crucial to enable research and development (R&D) and the commercialization of AI. However, policymakers should be aware of the dynamic limitations of statutory exceptions in copyright law, stemming from strategic behavior of rights holders in response to those exceptions. We discuss in detail how a balanced approach, where mandatory licensing enables data access while compensating rights holders, may be more efficient in balancing intellectual property protection with the goal of enabling and promoting AI innovation.

2. Copyright and Training Data

Copyright and neighboring rights play an important role in the data economy, particularly for ML and AI applications. However, the legal situation is complex and not at all harmonized across countries (Flynn et al., 2022). A detailed discussion of copyright law in the context of data as an input to ML and AI applications is provided in Kretschmer et al. (2024), here we focus only on key aspects that are most relevant for our empirical analysis below.

First, AI training data can be simply facts or sensor readings or include works dedicated to the public domain or works that are out-of-copyright. However, it can also include could include works protected by copyright (text, audiovisual content). Copyright law grants authors several exclusive rights over their works, including the right to reproduce, distribute, create derivative works, and publicly perform the work. In the process of collecting and processing data, any digital copy of copyrighted material may infringe copyright, in particular the right of reproduction. In addition, the processing of datasets, as is often necessary for ML, can amount to an adaptation within the scope of the exclusive right (e.g., cropping or rotating an image or adjusting its color scale). Second, some countries also protect datasets (i.e., the collection of datapoints) by copyright or via sui generis rights. For example, in the EU, Directive 96/9/EC offers intellectual property protection specifically for databases that, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation. Such protection aligns with traditional intellectual property rights, ensuring that the creative aspects of database design are safeguarded under existing copyright frameworks. Simultaneously, the directive introduces sui generis protection, which is distinct from copyright law, aimed at protecting the substantial investment in obtaining, verifying, or presenting the contents of the database, even if the database does not qualify for copyright protection. In contrast, the protection of databases in the US relies primarily on contract law or traditional copyright law, which only covers databases that exhibit a sufficient degree of creativity in their selection or arrangement of content, leaving purely factual compilations without substantial creative effort outside

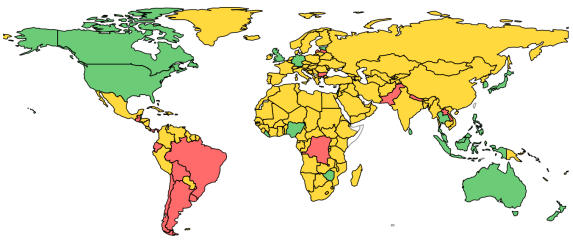

Figure 1. Map of copyright exception regimes

Note: The maps depict countries with copyright exceptions (*green*) relevant for AI, countries with restricted copyright exceptions (*yellow*) and countries without copyright exceptions relevant for AI (*red*). Source: Fiil-Flynn et al. (2022).

the scope of copyright protection.

Statutory data access, through fair use and TDM exceptions in copyright law, in particular, can play a pivotal role in enabling the analysis and use of large datasets across diverse research fields (Filippov and Hofheinz, 2016; Palmedo, 2019; Handke et al., 2021). However, the collection and processing of data is heavily influenced by national copyright laws, which vary widely in their approach to relevant exceptions. While some jurisdictions offer broad exceptions that facilitate extensive data analysis and sharing, others impose stringent restrictions that can limit the scope and impact of research. Flynn et al. (2022) provide a dataset on copyright exceptions across more than 200 countries. Building on their work, Figure 1 depicts differences across copyright by categorizing national law according to the breadth of exceptions. The following color-coding has been applied to three categories of breadth of copyright exceptions.

Green: Jurisdictions in this category allow for the reproduction and sharing of data for all users and all works, with some restrictions on commercial uses.

Yellow: Jurisdictions in the yellow category generally allow TDM, but impose certain restrictions. In some countries, datasets can be reproduced without limitations on who can use them, often even for commercial purposes, but sharing the reproduced data is prohibited. In other countries, only individuals can reproduce entire works for private or non-commercial use but cannot share them. Some jurisdictions allow certain institutions, such as universities and research organizations, to make copies of entire works for research purposes. Finally, some countries impose specific limitations on the types of works that can be reproduced. For example, they may prohibit the reproduction of whole books but allow for the analysis of other types of content, like scientific articles or datasets from social media.

Red: The most restrictive category, where TDM is heavily limited. Only excerpts of works can be used for TDM purposes, significantly constraining the scope of analysis. This restriction can hinder comprehensive R&D that relies on large datasets.

The US is an example of a "green country". The legal framework governing the training of AI models in the US is significantly influenced by the fair use doctrine, which is evaluated based on four key factors: (1) the purpose and character of the use, (2) the nature of the copyrighted work, (3) the amount and substantiality of the portion used, and (4) the effect of the use on the potential market for or value of the copyrighted work. The first factor examines whether the use of data is for a commercial purpose or for nonprofit educational purposes. This factor is particularly crucial in determining whether the use is considered transformative. The fourth factor assesses whether the use of copyrighted material could substitute for the original work or diminish its value, which directly impacts the economic interests of the copyright holder. These two factors often carry significant weight in fair use analyses related to AI, but all four factors must be weighed together in a holistic analysis (Samuelson, 2023).

Some EU member states, such as Germany, are also classified as green, but the majority of EU countries are classified as yellow. The lack of harmonization in the EU is worth discussing in more detail. The Directive on Copyright in the Digital Single Market (CDSM), introduced to address the evolving digital landscape and its impact on copyright and related rights, came into effect on June 7, 2019. The directive introduces significant changes, including TDM exceptions in Articles 3 and 4, which facilitate research by permitting the extraction of information from large datasets without infringing on copyrights. The EU Member States were required to transpose the directive into their national laws by June 7, 2021. However, as of early 2023, full implementation remains inconsistent across the Union.[1] Only a few countries have successfully met the transposition deadline. Despite its objectives, the directive's complexity and the contentious nature of some provisions have led to delays and different implementation approaches among Member States. The intricate implementation process of the CDSM highlights the challenges in harmonizing copyright laws across the EU and reflects different national priorities and legal traditions. As Member States continue to work on integrating the directive into their national

[1] See https://www.create.ac.uk/cdsm-implementation-resource-page/ for a comprehensive overview of the implementation status.

frameworks, ongoing legal and regulatory adjustments are to be anticipated, with the potential for significant long-term impacts on the digital market and copyright enforcement in Europe. This is especially relevant because the EU's latest sui generis law to regulate AI makes a direct reference to compliance with EU law on copyright and related rights (see EU AI Act, Art. 53).

Finally, it is important to note that exceptions present in the copyright law of a range of countries, such as for research or TDM may not cover all of the activities in a typical workflow used by researchers or commercial players (Kretschmer et al., 2024). That is, even in a "green" country, there can still be substantial legal uncertainty about specific aspects in the ML value chain that are important for R&D as well as the commercialization of AI technologies.

3. Data and Methods

In this section, we introduce the data sources we tap to study the correlation between statutory access in copyright law and AI innovation. As discussed above, data from Fiil-Flynn et al. (2022) provides us with information about the breadth of exceptions in copyright law across more than 200 countries.

3.1. Research output: arXiv

To measure research output, we access fine-grained information about AI publications. For the entire corpus of about 379,105 papers listed on the Machine Learning community website Paperswithcode, published between April 2007 and March 2024, we downloaded the full text from the preprint server arXiv. By parsing the first page of the PDF files, we could successfully obtain the email address domain of the author(s) for 247,403 papers (65.3%). We then linked domain names to countries, using extensive lists of universities and research institutions and industry organizations as well as the country-specific top-level domain (TLD). After excluding email domains to which we cannot unambiguously assign a country of the email address owner (e.g., `gmail.com`), we were able to approximate countries of 1,148,202 author(s) of 215,448 papers (56.8%). The top ten countries make up 75% of the authors, with the United States (32.2%) leading China (16.6%), the United Kingdom (6.1%), Germany (4.9%), Canada (3.2%), Australia (2.8%), France (2.8%), Italy (2.3%), India (2.3%) and Japan (2.1%).

3.2. Open source code: GitHub

To obtain the universe of GitHub repositories (in short: repos) in the AI/ML space, we followed the methodology developed in Gonzalez et al. (2020), which is widely used, e.g. by the OECD AI Policy Observatory (`https://oecd.ai/`). For this, we make use of the fact that owners can tag their GitHub repos according to the contents with so-called topics. Using the extensive list of AI/ML topics from Gonzalez et al. (2020), we collected the list of corresponding repos using the GitHub API, which amounts to 13,740 repos that existed on GitHub as of June 2023. We then filtered out repos that contained only educational content (tutorials, courses, etc.) using a manually coded training dataset and basic natural language processing methods. With this clean list of AI repos, we obtained all push events of these repos from GHArchive and parsed the email address domain of all authors of commits within the event's payload. We then filtered for commits to the default branch, which resulted in a total number of 2,884,313 commits. We excluded commits with non-informative details, e.g., where the email domain was `users.noreply.github.com`, which GitHub uses to obscure email addresses according to users' privacy settings.[2] We then applied the same method of mapping email domains to countries as described in section 3.1. A majority of commits are made by contributors with email domains that cannot be unambiguously mapped to a country, such as `.com` and `.ai`. In our final sample, we can approximate the country of 12,897 contributors who made a total of 359,094 (12.4%) commits to 3,945 repos (28.7%) between October 2011 and May 2024.

The distribution of contributor countries is skewed. The top ten countries make up 82.3% of the contributions in our final sample, with the United States (24.2%) leading Germany (14.7%), Russia (14.6%), Switzerland (5.9%), Japan (5.4%), France (5.2%), the United Kingdom (4.4%), China (3.4%), Norway (2.6%) and India (1.9%).

3.3. AI Patents: USPTO

In addition to software-heavy AI applications that we can measure with GitHub data, we turn to the AI Patent Dataset (AIPD) developed by Giczy et al. (2022) to also measure more hardware-based AI applications. In addition, patent data can also serve as a proxy for the commercialization of technology, as patent protection can attract venture capital funding and allows for the appropriation of rents in a market (Conti et al., 2013). AIPD classifies patents from the United States Patent and Trademark Office (USPTO) into AI and non-AI related technologies. Giczy et al. (2022) defined AI through eight key component technologies: knowledge processing, speech, AI hardware, evolutionary computation, natural language

[2] See `https://github.com/orgs/community/discussions/41101`.

processing, machine learning, computer vision, and planning/control. These components were employed to identify relevant patent documents while acknowledging potential overlaps between technologies. For instance, natural language processing might utilize underlying machine learning methods. Each AI component was defined in detail, such as knowledge processing involving methods to derive new facts from a knowledge base, and AI hardware encompassing physical components. This comprehensive approach ensures a thorough identification of AI-related patents.

For the purposes of our analysis, however, it is crucial to make cross-country comparisons. For this, we can exploit the fact that the USPTO accepts applications from non-US entities.[3] However, of course, this comes with limitations. The analysis can only reveal a lower bound since we can only capture patents from international inventors issued in the US. However, given that the US is an important market (both for consumer- and business-facing applications), we expect that inventors/firms would apply for patents in their home countries and the US, at least with respect to their most important inventions.

We observe a total of 627,115 AI patents issued to non-US inventors between January 1980 and December 2020. The top ten countries make up 81% of the contributions in our final sample, with Japan (22.5%) leading Germany (9.7%), South Korea (7.9%), China (7.7%), Canada (7.4%), India (6.9%), Israel (5.9%), the United Kingdom (5.9%), France (4.1%) and Taiwan (3.2%).

3.4. Ventures: Crunchbase

To measure commercialization and professionalization, we use data on the founding of new organizations (short: ventures). Following related literature (e.g. Montanaro et al., 2024), we access data from Crunchbase, a comprehensive platform that provides detailed information about high-tech organizations (more than 90% companies, but also non-profit organizations). The platform aggregates data from various sources such as organization websites, news articles, and user-generated content. Key variables in the Crunchbase dataset include organization name, industry, headquarters location, founding date, funding details, and key personnel. The dataset encompasses organizations from a diverse range of countries, thus providing a global perspective on the entrepreneurial landscape.[4] Organizations in Crunchbase are categorized into numerous industry groups, such as technology, healthcare, finance, and more, allowing for granular analysis within specific sectors. To define a measure of the number of AI ventures per country and month, we filtered the dataset by the industry categories that included the terms "Artificial Intelligence" and "Data" and then count the number of organizations listed for the first time on Crunchbase in each country and month.

Using this definition, we observe a total of 152,797 AI ventures founded between January 2008 and December 2023 in the countries for which we have information about fair use or TDM exceptions in copyright law. About 74% of the ventures are located within ten countries. The clear leader is the United States (38.7%), followed by the United Kingdom (7.3%), India (6.6%), Germany (4.3%), Canada (7.4%), China (2.9%), Brazil (2.9%), France (2.8%), Japan (2.5%) and Italy (2.4%).

3.5. Econometric specification

To estimate correlations between statutory data access and AI innovation across countries, we run two simple linear regression models. We can distinguish between level effects (i.e., estimate an intercept parameter for different TDM regimes) and dynamic effects (i.e., estimate a slope parameter for different TDM regimes).

The model for level effects is defined as: $\text{Log}(\text{Y}_{it}^{k} + 1) = \alpha_1 \textit{Green}_i + \alpha_2 \textit{Yellow}_i + \delta Control_i + \varepsilon_{it}$,

and model for dynamic effects is defined as: $\text{Log}(\text{Y}_{it}^{k} + 1) = \alpha_1 \textit{Green}_i + \alpha_2 \textit{Yellow}_i + \beta_1 \left(Trend_t^k \times \textit{Green}_i\right) + \beta_2 \left(Trend_t^k \times \textit{Yellow}_i\right) + \gamma Trend_t^k + \delta Control_i + \varepsilon_{it}$,

where Y_{it}^k is the count of outcome k (Papers, Repos, Patents, Ventures) in country i in month t. $\textit{Green}_i$ and $\textit{Yellow}_i$ are dummy variables indicating green and yellow TDM exception status, respectively. $Trend_t^k$ is a linear time trend, normalized to the first month of the respective k dataset. Country-specific control variables do not vary over time.

4. Results

4.1. Country characteristics and copyright exceptions

As a first step, we characterize groups of countries within the three categories of copyright exceptions (green, yellow, red). Table 1 relates country

[3] Ideally, we would like to analyze global patent data and study applications by domestic firms with their domestic authorities. To be best of our knowledge, however, data with comparable levels of detail is not available internationally.

[4] However, it is not clear whether Crunchbase data is representative of a country's founding activities. While Crunchbase is used frequently in Entrepreneurship and Innovation research, we are not aware of academic work that assesses the national representativeness, perhaps because often there is no census or not a sufficiently detailed or up-to-date census of a country's economy. This leaves a potential caveat for our study. Given that Crunchbase is located in the US, it is likely that the coverage of North American-based organizations is more accurate.

Table 1. Country characteristics

	Green vs. All		Green vs. Red		Green vs. Yellow	
	(1)	(2)	(3)	(4)	(5)	(6)
OECD	0.0448	-0.1779*	0.2084	-0.1166	0.0301	-0.2141*
	(0.09113)	(0.09988)	(0.23185)	(0.21189)	(0.09995)	(0.11648)
Log(GDPperCapita)	0.0471**	0.0498*	0.0382	0.0929	0.0526**	0.0590**
	(0.02283)	(0.02544)	(0.07555)	(0.08101)	(0.02471)	(0.02865)
Log(Population)	0.0190	0.0299	-0.0180	0.0314	0.0220	0.0358
	(0.01511)	(0.02149)	(0.03326)	(0.05495)	(0.01614)	(0.02354)
R&D Share		0.1314***		0.1462**		0.1305***
		(0.04295)		(0.06794)		(0.04660)
Observations	195	145	43	37	180	130

Note: The dependent variable is an indicator of whether a country falls in the green category. In columns (1) and (2), the comparison group includes yellow and red countries, in columns (3) and (4) it only includes red countries, and in columns (5) and (6), it includes yellow countries. White-robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$

characteristics to the breadth of exceptions in national copyright law.[5] In column (1), we compare green countries to all other countries. The regression suggests that countries with higher income (higher Gross Domestic Product, GDP per capita) are more likely to have broad copyright exceptions, but we do not see correlations with population size or a country's membership in the OECD. In column (2) we additionally add information on a country's overall investment in R&D as a share of the GDP. This information is only available for a subset of countries. The estimates suggest that countries with more R&D investment are more likely to offer broad exceptions in their copyright law. After controlling for R&D investment, the coefficient of the indicator for a country's membership in the OECD turns negative. In columns (3) and (4), we compare countries with green TDM status to countries with red TDM status. Except that green countries tend to have higher investments into R&D, we do not find significant differences in terms of income, population and OECD membership. Turning to columns (5) and (6), it becomes clear that green countries differ significantly from yellow countries. The results from columns (1) and (2) seem to be driven by yellow countries, not red countries. In sum, this exercise suggests that richer countries with higher R&D spending are more likely to have broader copyright exceptions.

4.1.1. Model-free evidence

We now turn to model-free evidence on the relationship between exceptions in copyright law and our four measures of AI innovation. Figure 4.1.1 shows that countries with broad copyright exceptions (green countries) tend to publish more ML papers, commit more to AI repos, file for more AI patents and found more AI organizations. The difference between green, yellow, and red countries is not only visible in levels but also in the rate of increase in the late 2010s and 2020s. However, as discussed above, green countries are also typically richer and tend to invest more in R&D. Hence, these factors may drive the dynamics depicted in Figure 4.1.1 and not (or to a much smaller scale) the presence or absence of broad exceptions in copyright law. To investigate this further, we rely on a regression analysis where we can control for income, population, and R&D investment at the country level.[6]

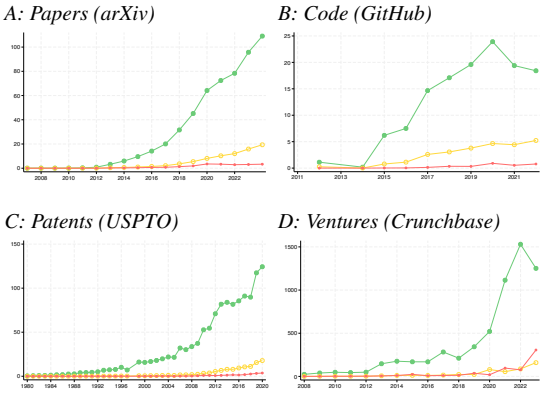

Figure 2. AI innovation measures, group averages

Note: Solid lines indicate the average for countries with copyright exceptions (*green*), solid lines with hollow circles indicate the average for countries with restricted copyright exceptions (*yellow*), and dashed lines indicate the average for countries without copyright exceptions (*red*).

4.1.2. Regression results

In Table 2, we look at the monthly number of academic papers published on arXiv and the monthly number of contributions to AI repos on GitHub. The coefficient estimate of *TDM: Green* in column (1), without any control variables, suggests that green countries see 76% more papers (=exp(0.5634)-1) than red countries. The coefficient of *TDM: Green* is reduced by almost a factor of 5 after we add the control variables in column (2) and becomes insignificant. We get similar results when focusing on commits to AI repositories on GitHub in column (4). The estimated coefficient of *TDM: Green* suggests that users with email addresses from green countries make 69% more contributions than users from red countries. In column (5), where we add country-specific controls, the coefficient of *TDM: Green* is more than 6 times smaller and not significant. Overall, these results suggest that it is difficult to attribute the difference in levels of paper publications and

[5] Country-level economic indicators come from the Worldbank's Open Data Platform but are not available for all countries for which we have information on copyright law. Given that our dataset on copyright law from Fiil-Flynn et al. (2022) does have time-variant information, we aggregate country characteristics. In particular, we take the maximum value available between 1980 and 2022.

[6] Note that results remain almost identical when we control for nominal GDP rather than GDP per capita and population separately.

Table 2. Papers and Code

	Papers			Repos		
	(1)	(2)	(3)	(4)	(5)	(6)
TDM: Green	0.5634***	0.1213	-0.6737***	0.5241***	0.0834	-0.3103*
	(0.16943)	(0.09255)	(0.18396)	(0.17837)	(0.11509)	(0.15650)
TDM: Yellow	0.0907	0.0921**	-0.0868	0.1641**	0.0754	-0.0477
	(0.07639)	(0.04063)	(0.11846)	(0.06280)	(0.06776)	(0.11564)
R&D Share		0.1809***	0.1809***		0.1992***	0.1992***
		(0.04153)	(0.04153)		(0.07330)	(0.07330)
Log(GDPperCapita)		0.1719***	0.1719***		0.1280**	0.1280**
		(0.02925)	(0.02925)		(0.04965)	(0.04966)
Log(Population)		0.1740***	0.1740***		0.1455***	0.1455***
		(0.02754)	(0.02754)		(0.03101)	(0.03101)
TimeTrend			0.0026***			0.0016*
			(0.00100)			(0.00095)
TimeTrend × TDM: Green			0.0078***			0.0049**
			(0.00225)			(0.00203)
TimeTrend × TDM: Yellow			0.0017			0.0015
			(0.00120)			(0.00116)
Mean DV	0.1684	0.1684	0.1684	0.0834	0.0834	0.0834
Observations	149,480	126,250	126,250	59,000	48,500	48,500

Note: The dependent variable in columns (1)–(3) is the log(+1) monthly number of papers from authors in countries with the respective TDM regime. The dependent variable in columns (4)–(6) is the log(+1) monthly number of contributions (=commits in push events) from users in countries with the respective TDM regime. For all columns, the data is at the country-month-usertype-level, where the user type is either company, university, or individual as inferred from the email address of a user. Standard errors clustered at the country level in parentheses. $^{*}\ p < 0.10, ^{**}\ p < 0.05\ ^{***}\ p < 0.01$

GitHub contributions between green and red countries to differences in copyright law. As a next step, we investigate dynamic differences, taking the growth patterns shown in Figure 4.1.1 into consideration. We model this with a simple time trend, normalized to the first observed time period, in columns (3) and (6). Note that this model specification allows us to specifically capture and quantify the growth in publications and contributions, while the control variables condition on factors that affect the level of publications and contributions. The estimated coefficients of *TimeTrend* × *TDM: Green* imply that every month, the number of publications and contributions in green countries increases about 0.5%-1% quicker than the number of publications and contributions in red countries. This reflects the steeper slope of the green curve compared to the red curve in Panels A and B of Figure 4.1.1.

We repeat the same analysis with respect to patents and newly founded ventures. The results in columns (1) and (3) of Table 3 imply that green countries apply for 152% more patents and see 96% more new ventures than red countries. After controlling for observable country characteristics in columns (2) and (5), the significant coefficients of *TDM: Green* imply a difference of 38% in terms of patents and 32% in terms of new ventures. Turning to dynamics in columns (3) and (6), the estimated coefficients of *TimeTrend* × *TDM: Green* imply that every month, the number of patents and new ventures in green countries increases about 0.5% quicker than the number of patents and new ventures in red countries.

Table 3. Patents and Ventures

	Patents			Ventures		
	(1)	(2)	(3)	(4)	(5)	(6)
TDM: Green	0.9206***	0.3232*	-0.8473***	0.7230***	0.3134**	-0.2722
	(0.27562)	(0.18299)	(0.28184)	(0.25411)	(0.14954)	(0.17979)
TDM: Yellow	0.1024	0.0283	-0.2219	-0.0423	0.0126	0.0515
	(0.07087)	(0.06644)	(0.16833)	(0.11296)	(0.06094)	(0.12833)
R&D Share		0.4535***	0.4535***		0.2482***	0.2482***
		(0.05198)	(0.05198)		(0.05927)	(0.05927)
Log(GDPperCapita)		0.1475***	0.1475***		0.2092***	0.2092***
		(0.03510)	(0.03510)		(0.03646)	(0.03646)
Log(Population)		0.1650***	0.1650***		0.2208***	0.2208***
		(0.03280)	(0.03280)		(0.03709)	(0.03709)
TimeTrend			0.0011**			0.0050***
			(0.00048)			(0.00145)
TimeTrend × TDM: Green			0.0048***			0.0061***
			(0.00132)			(0.00190)
TimeTrend × TDM: Yellow			0.0010*			-0.0004
			(0.00059)			(0.00155)
Mean DV	0.1355	0.1355	0.1355	0.3566	0.3566	0.3566
Observations	94,464	67,896	67,896	37,248	27,456	27,456

Note: The dependent variable in columns (1)–(3) is the log(+1) monthly number of AI patents issued by the USPTO from inventors in countries with the respective TDM regime. We exclude inventors from the United States. The dependent variable in columns (4)–(6) is the log(+1) monthly number of newly founded ventures in countries with the respective TDM regime. Standard errors clustered at the country level in parentheses. $^{*}\ p < 0.10, ^{**}\ p < 0.05\ ^{***}\ p < 0.01$

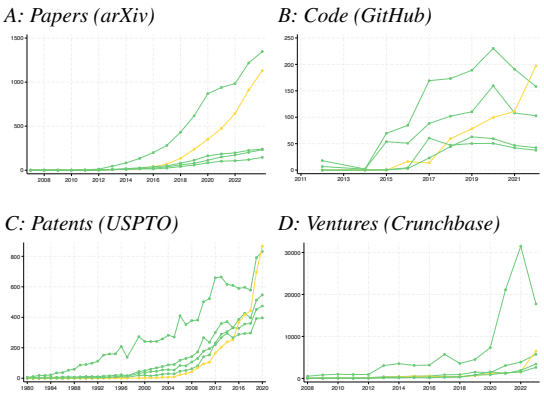

Figure 3. AI innovation measures, top 5 countries

Note: For each measure, the respective 5 top countries are depicted. Colors indicate the copyright exception classification for a country.

Taken together, these results suggest that for research and open source software development, there is no level effect between green and red or yellow and red countries that cannot be explained by the country's overall levels of R&D investment, aggregate income and size. However, we see a faster pace of research and open source software development in countries that make access to data easier with more liberal copyright laws that provide unrestricted exceptions for TDM.

4.2. Discussion of internal validity

4.2.1. The role of leading countries

Given the skewed distribution of the innovation outcomes, where some countries are clear leaders, it seems likely that the results are driven by these outliers. Zooming into the top five countries in every

Table 4. Papers and Code, excluding top 5

	Papers			Repos		
	(1)	(2)	(3)	(4)	(5)	(6)
TDM: Green	0.2954**	-0.0006	-0.4578***	0.2469**	-0.0257	-0.2117*
	(0.12531)	(0.07340)	(0.16349)	(0.10483)	(0.08123)	(0.12099)
TDM: Yellow	0.0731	0.0779*	-0.0719	0.1297**	0.0715	-0.0041
	(0.07451)	(0.04069)	(0.10337)	(0.05875)	(0.05121)	(0.08974)
R&D Share		0.1565***	0.1565***		0.1566***	0.1566***
		(0.02609)	(0.02609)		(0.04878)	(0.04878)
Log(GDPperCapita)		0.1485***	0.1485***		0.0935**	0.0935**
		(0.02526)	(0.02526)		(0.03754)	(0.03754)
Log(Population)		0.1374***	0.1374***		0.1027***	0.1027***
		(0.02492)	(0.02492)		(0.02639)	(0.02639)
TimeTrend			0.0026***			0.0016*
			(0.00100)			(0.00095)
TimeTrend × TDM: Green			0.0045**			0.0023
			(0.00190)			(0.00158)
TimeTrend × TDM: Yellow			0.0015			0.0009
			(0.00117)			(0.00110)
Mean DV		0.1684	0.1684	0.0834	0.0834	0.0834
Observations	144,430	121,200	121,200	56,500	46,000	46,000

Note: The dependent variable in columns (1)–(3) is the log(+1) monthly number of papers from authors in countries with the respective TDM regime. The dependent variable in columns (4)–(6) is the log(+1) monthly number of contributions (=commits in push events) from users in countries with the respective TDM regime. For all columns, the data is at the country-month-usertype-level, where the user type is either company, university, or individual as inferred from the email address of a user. Standard errors clustered at the country level in parentheses. $^{*}\ p < 0.10, ^{**}\ p < 0.05\ ^{***}\ p < 0.01$

Table 5. Patents and Ventures, excluding top 5

	Patents			Ventures		
	(1)	(2)	(3)	(4)	(5)	(6)
TDM: Green	0.4740**	0.0622	-0.7183**	0.3119*	0.0395	-0.4127***
	(0.19772)	(0.11134)	(0.27975)	(0.16209)	(0.08696)	(0.15065)
TDM: Yellow	0.0896	0.0371	-0.1809	-0.0600	-0.0303	0.0248
	(0.06988)	(0.05452)	(0.14944)	(0.11165)	(0.06013)	(0.10494)
R&D Share		0.4088***	0.4088***		0.2225***	0.2225***
		(0.04042)	(0.04042)		(0.03626)	(0.03626)
Log(GDPperCapita)		0.1228***	0.1228***		0.1749***	0.1749***
		(0.03075)	(0.03075)		(0.02773)	(0.02773)
Log(Population)		0.1291***	0.1291***		0.1540***	0.1540***
		(0.03073)	(0.03073)		(0.01920)	(0.01921)
TimeTrend			0.0011**			0.0050***
			(0.00048)			(0.00145)
TimeTrend × TDM: Green			0.0032**			0.0047**
			(0.00130)			(0.00192)
TimeTrend × TDM: Yellow			0.0009			-0.0006
			(0.00058)			(0.00155)
Mean DV	0.1355	0.1355	0.1355	0.3566	0.3566	0.3566
Observations	92,004	65,436	65,436	36,288	26,496	26,496

Note: The dependent variable in columns (1)–(3) is the log(+1) monthly number of AI patents issued by the USPTO from inventors in countries with the respective TDM regime. We exclude inventors from the United States. The dependent variable in columns (4)–(6) is the log(+1) monthly number of newly founded ventures in countries with the respective TDM regime. Standard errors clustered at the country level in parentheses. $^{*}\ p < 0.10, ^{**}\ p < 0.05\ ^{***}\ p < 0.01$

AI innovation measure in Figure 4.2.1, we see that most but not all of the top countries have a copyright regime with broad exceptions (green).

We repeat the analysis and exclude the top five countries (with respect to each outcome variable) from the sample. The results reported in Tables 4 and 5 are qualitatively the same compared to those reported in Tables 2 and 3, with estimated coefficients of *TDM: Green* and *TimeTrend* × *TDM: Green* being smaller in size but not statistically different in the sense that 90% confidence bands do not overlap.[7]

4.2.2. Causality

A key caveat of our study lies in its correlational nature. Unobserved country- and time-specific factors might contribute to the patterns revealed in our correlational analysis. The ideal setting would involve a randomized experiment that assigns a certain copyright status to some actors and not to others. Clearly, it is not possible to experimentally vary copyright law at the individual or national levels. Future work could take advantage of law changes, e.g., after the EU's Directive on Copyright in the Digital Single Market is finally implemented in the national law of the member states. Such variation would further strengthen the statistical analysis because a difference-in-differences estimation with fixed effects would remove unobserved and time-invariant country-specific variation. Even this approach would not be entirely clean. In the interconnected digital world, there might still be time-variant unobserved factors determining the change in copyright law as well as innovation in AI, e.g., coming from industrial policy, competition policy, or privacy law goals. Further, there might be spillover effects of copyright policy in other jurisdictions. In sum, while we need to acknowledge the lack of causality in our estimates, we hope to still provide informative evidence that should be useful to guide policy discussions, at least more useful than not having any evidence on a pressing policy issue.

5. Policy implications

Our results suggest that statutory access to data through exceptions in copyright law is positively correlated with AI innovation. We find that countries with broad copyright exceptions tend to see faster growth in research output, software development efforts and more and faster growth in patents and ventures focusing on AI technology.

What are the potential downsides of broad copyright exceptions for AI training data? Recent evidence suggests important dynamic effects when individual copyright holders strategically limit access to existing works or decide which types of new works to create or whether to create any at all. Peukert et al. (2024) study contributors on Unsplash, a popular stock image platform that launched an AI research program by releasing a dataset of 25,000 randomly selected images for commercial use. The study finds that affected contributors left the platform at a higher-than-usual rate.

[7] The lower end of the 90% confidence band of the point estimate of *TDM: Green* in column 1 of Table 2 is 0.2313, which is smaller than 0.2954, the point estimate of *TDM: Green* in column 1 of Table 4, and so on.

The remaining contributors substantially slowed down the rate of new uploads and also changed the variety and novelty of contributions to the platform and, therefore, to future training datasets. As a result of smaller and strategically selected training datasets, the quality of AI output may degrade in the long run.

This implies that broad copyright exceptions can work if only the existing stock of works is important for AI innovation, but strategic behavior may impact the flow of new works and, therefore, the flow of training data in the future. Because of a similar mechanism of strategic behavior, restrictive exceptions that explicitly give decision power to rights holders are problematic. Consider the example of the EU Copyright directive where TDM is allowed unless copyright holders opt out, demanding license payments. In a conceptual paper, Martens (2024) argues that opt-out of TDM can be seen as economically inefficient overprotection of copyright, as free use of media for AI training does not impact media sales to consumers but merely strengthens copyright holders' bargaining positions, leading to windfall profits without benefiting consumer surplus or social welfare. An opt-out regime reduces the quantity and quality of the stock of training data, increases transaction costs, hinders competition, and can, therefore, slow down innovation. Such considerations seem to be backed by the correlational evidence we provide, where countries without or with restrictive copyright exceptions tend to fall behind those with broad exceptions – either in trends or both levels and trends across our four measures of R&D and innovation.

Further, if copyright holders can (strategically) restrict access to training data, this can lead to bias in the data available to AI models. As highlighted by Levendowski (2018), exclusive rights enabled by copyright law can restrict bias mitigation techniques, including testing AI through reverse engineering, implementing algorithmic accountability processes, and market competition. Further, AI developers may rely more on easily accessible, low-risk data sources, even when such data is known to be biased. However, broad statutory exceptions, such as fair use, are capable of addressing concerns about AI bias, largely because the normative principles underlying fair use align with the objectives of mitigating AI bias and contributing to the development of fairer AI systems. This is important to highlight given that this implies a difficult tradeoff for European policy. For example, the European Union (EU) AI Act reinforces copyright by requiring providers of general-purpose AI models to put in place policies to respect EU copyright law (Art. 53), but also that "Training, validation and testing datasets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose." (Art. 10(3)).

A potentially economically more efficient system without detrimental dynamic effects could be mandatory licensing (Geiger and Iaia, 2024). Mandatory licensing offers a pragmatic solution to the trade-off between preserving the dynamic incentives of creators and facilitating the continued investment in AI technology that relies on access to copyright-protected data. By implementing a mandatory licensing framework, creators are assured compensation for the use of their works, thereby maintaining their motivation to generate new content. At the same time, AI developers gain predictable access to the stock of copyrighted works as training data as well as to a flow of copyrighted works that is not affected by strategic behavior.

A well-designed mandatory licensing model could incorporate modern technology for tracking usage and distributing royalties, thus addressing the administrative burdens seen in similar existing systems. However, given the international scale and the multi-modal nature of training data, a few important issues remain to be solved. For example, who would be responsible for overseeing licensing? Would it be managed on a country-by-country basis, or would it fall under the jurisdiction of a global nonprofit organization? Should different types of content have their own separate licensing bodies? How could we prevent potential misuse of the system? Would there be requirements for transparency? And what role, if any, would the government play in this process? Further, and quite fundamentally, it would be important to clarify what constitutes use – a broader set of tools or research program, or a model, or even just the fine-tuning of a model? By providing a clear, structured, and equitable framework, mandatory licensing for AI could ensure that creators receive fair compensation while enabling AI developers to dynamically access the necessary data.

6. Conclusions

This study set out to explore the relationship between statutory access to data through copyright exceptions and innovation in AI, asking: What is the relationship between statutory access to data and innovation in AI?

We show that broad copyright exceptions, particularly those enabling access to data TDM and fair use, are positively correlated with higher levels of AI innovation. Jurisdictions with broader copyright exceptions tend to experience faster growth in AI-related research output, software development, patent filings, and venture formation. However, it is important to note that we cannot definitively establish a

causal link between copyright law and AI innovation. Despite this limitation, our study provides the best available evidence to date on the relationship between copyright exceptions and AI innovation.

Our findings suggest that facilitating statutory access to data is crucial for fostering innovation in AI. However, the potential for strategic behavior by rights holders, which could limit access to new works and impact the quality of AI training data, poses a significant challenge. While broad copyright exceptions may support current AI innovation by allowing the use of existing works, they might also inadvertently discourage the creation of new works, leading to long-term consequences for the availability and (broadly defined) quality of training data.

To address these challenges, we propose that policymakers should consider the adoption of a mandatory licensing framework to balance the dynamic interests of both copyright holders and AI developers. Such a system could provide access to copyrighted works for AI training while ensuring that creators receive fair compensation, thereby maintaining incentives for the continued creation of new content. Further research and policy development will be necessary to refine such an approach and ensure that it effectively meets the needs of both the AI industry and rightsholders.

References

Bessen, J., Impink, S. M., Reichensperger, L., and Seamans, R. (2022). "The role of data for AI startup growth." *Research Policy*, *51*(5), 104513.

Bianchini, S., Müller, M., and Pelletier, P. (2022). "Artificial intelligence in science: An emerging general method of invention." *Research Policy*, *51*(10), 104604.

Conti, A., Thursby, J., and Thursby, M. (2013). "Patents as signals for startup financing." *The Journal of Industrial Economics*, *61*(3), 592–622.

Fiil-Flynn, S. M., Butler, B., Carroll, M., Cohen-Sasson, O., Craig, C., Guibault, L., Jaszi, P., Jütte, B. J., Katz, A., Quintais, J. P., et al. (2022). "Legal reform to enhance global text and data mining research." *Science*, *378*(6623), 951–953.

Filippov, S., and Hofheinz, P. (2016). "Text and Data Mining for Research and Innovation." *Lisbon Council Interactive Policy Brief*, *23*, 3.

Flynn, S., Schirru, L., Palmedo, M., and Izquierdo, A. (2022). "Research exceptions in comparative copyright." *PIJIP/TLS Research Paper Series*, *75*.

Geiger, C., and Iaia, V. (2024). "The forgotten creator: Towards a statutory remuneration right for machine learning of generative ai." *Computer Law & Security Review*, *52*, 105925.

Giczy, A. V., Pairolero, N. A., and Toole, A. A. (2022). "Identifying artificial intelligence (AI) invention: A novel AI patent dataset." *The Journal of Technology Transfer*, *47*(2), 476–505.

Gonzalez, D., Zimmermann, T., and Nagappan, N. (2020). "The state of the ML-universe: 10 years of Artificial Intelligence & Machine Learning software development on Github." In *Proceedings of the 17th International conference on mining software repositories*, 431–442.

Handke, C., Guibault, L., and Vallbé, J.-J. (2021). "Copyright's impact on data mining in academic research." *Managerial and Decision Economics*, *42*(8), 1999–2016.

Kretschmer, M., Margoni, T., and Oruc, P. (2024). "Copyright law and the lifecycle of machine learning models." *IIC-International Review of Intellectual Property and Competition Law*, 1–29.

Levendowski, A. (2018). "How copyright law can fix artificial intelligence's implicit bias problem." *Wash. L. Rev.*, *93*, 579.

Martens, B. (2024). "Economic arguments in favour of reducing copyright protection for generative ai inputs and outputs." Tech. rep., Bruegel.

Montanaro, B., Croce, A., and Ughetto, E. (2024). "Venture capital investments in artificial intelligence." *Journal of Evolutionary Economics*, 1–28.

Palmedo, M. (2019). "The impact of copyright exceptions for researchers on scholarly output." *Efil Journal of Economic Research*.

Peukert, C., Abeillon, F., Haese, J., Kaiser, F., and Staub, A. (2024). "Strategic behavior and ai training data." *arXiv preprint arXiv:2404.18445*.

Peukert, C., and Windisch, M. (2024). "The economics of copyright in the digital age." *Journal of Economic Surveys*, *forthcoming*.

Samuelson, P. (2023). "Generative AI meets copyright." *Science*, *381*(6654), 158–161.

Wu, L., Hitt, L., and Lou, B. (2020). "Data analytics, innovation, and firm productivity." *Management Science*, *66*(5), 2017–2039.