IIC (2024) 55:110–138
https://doi.org/10.1007/s40319-023-01419-3

ARTICLE

Copyright Law and the Lifecycle of Machine Learning Models

Martin Kretschmer · Thomas Margoni · Pinar Oruç

Accepted: 11 December 2023 / Published online: 1 February 2024

Abstract Machine learning, a subfield of artificial intelligence (AI), relies on large corpora of data as input for learning algorithms, resulting in trained models that can perform a variety of tasks. While data or information are not subject matter within copyright law, almost all materials used to construct corpora for machine learning are protected by copyright law: texts, images, videos, and so on. There are global policy moves to address the copyright implications of machine learning, in particular in the context of so-called "foundation models" that underpin generative AI. This paper takes a step back, exploring empirically three technological settings through detailed case studies. We set out the established industry methodology of a lifecycle of AI (collecting data, organising data, model training, model operation) to arrive at descriptions suitable for legal analysis. This will allow an assessment of the challenges for a harmonisation of rights, exceptions and disclosure under EU copyright law. The three case studies are:

The research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870626870626 (*reCreating Europe*: Rethinking digital copyright law for a culturally diverse, accessible, creative Europe). Case study 1 was developed with ESRC support for the Urban Big Data Centre (ES/L011921/1). Pinar Oruç prepared a first draft of the case studies as a postdoctoral researcher at CREATe, University of Glasgow.

M. Kretschmer (✉)
Professor of Intellectual Property Law, School of Law, University of Glasgow; Director of CREATe (UK Copyright and Creative Economy Centre), Glasgow, UK
e-mail: martin.kretschmer@glasgow.ac.uk

T. Margoni
Research Professor of Intellectual Property Law, Centre for IT and IP Law (CiTiP), Faculty of Law and Criminology, University of Leuven (KU Leuven), Leuven, Belgium
e-mail: thomas.margoni@kuleuven.be

P. Oruç
Lecturer in Commercial Law, University of Manchester, Manchester, UK
e-mail: pinar.oruc@manchester.ac.uk

1. Machine learning for scientific purposes, in the context of a study of regional short-term letting markets;
2. Natural Language Processing (NLP), in the context of large language models;
3. Computer vision, in the context of content moderation of images.

We find that the nature and quality of data corpora at the input stage is central to the lifecycle of machine learning. Because of the uncertain legal status of data collection and processing, combined with the competitive advantage gained by firms not disclosing technological advances, the inputs of the models deployed are often unknown. Moreover, the "lawful access" requirement of the EU exception for text and data mining may turn the exception into a decision by rightholders to allow machine learning in the context of their decision to allow access. We assess policy interventions at EU level, seeking to clarify the legal status of input data via copyright exceptions, opt-outs or the forced disclosure of copyright materials. We find that the likely result is a fully copyright-licensed environment of machine learning that may have problematic effects for the structure of industry, innovation and scientific research.

Keywords Copyright · Artificial intelligence · Text mining · Data mining · EU · Digital single market

1 Introduction

New data analysis methods are attracting global attention. Machine learning, a subfield of Artificial Intelligence (AI), is seen as a critical technology, in which algorithms are trained on data to recognise and predict patterns. Data scraping, the acquiring and structuring of information from online sources, is a typical first step for machine learning. The technologies of scraping, mining and learning are often conflated, as are the legal regimes under which they are regulated. The legal issues involved in the governance of data range from proprietary approaches (copyright, database rights) to privacy and data protection laws, and wider provisions on data access and data sharing (for example under competition law or data legislation).[1]

Copyright law has a direct impact on the processes of data scraping, mining and learning. What are known as "corpora", i.e. collections of information needed for training purposes, could include works protected by copyright, other related subject matter, or simply facts and data. When copyright or a related right are present, any digital copy, temporary or permanent, in whole or in part, direct or indirect, has the potential to infringe that right, in particular the economic right of reproduction. Furthermore, the changes made in the collected material can amount to an "adaptation" within the scope of the exclusive right. Relevant exceptions, such as for research or text and data mining, might not cover all the activities of researchers and firms in this area.

The layered protection for data is confusing for users and regulators. The technology of machine learning has developed in a legally grey zone, relying on an underlying lifecycle of data processing and analysis that has been established for

[1] Margoni and Kretschmer (2022); Margoni et al. (2023); Eben et al. (2023).

many years. Powerful models have already been trained and are being integrated in many services, at a rapidly accelerating pace. The arrival of generative AI, and the associated visibility of consumer-facing applications, has led to a string of lawsuits and proposed interventions testing the proprietary assumption about data inputs.[2]

In this paper, we aim to understand how machine learning technology developed as a set of legally relevant facts and analyse the implications of copyright interventions, such as exceptions, opt-outs or the forced disclosure of copyright materials. The results of three empirical case studies will aid legal classification and assessment of the relevant regulatory framework, focusing on the EU.

2 Methodology

Legal research of emerging technologies typically starts with an identification of a relevant legal domain and proceeds to a doctrinal analysis of the scope of specific concepts and rules. The analysis is then evaluated against practical implications, often using particular factual constructions (scenarios) to illuminate the potential effects of interpretations or interventions.

There are dangers inherent in this legal approach to policy making. The analysis often lags behind technological developments. Scenarios may be filtered via professional representations or trade bodies that were constituted in a different context, perpetuating past discussion. In a wider sense, policy making may be anecdotally driven by examples that surface through lobbying processes or the latest technological applications.

In the current policy context, the dominant scenarios derive from advances in so-called generative artificial intelligence (AI) which has become more visible with the release of user-facing applications, such as large language models (accessed e.g. via OpenAI's Chat-GPT) or generative image applications (such as Midjourney).

The research design informing this paper takes a more long-term perspective. Machine learning techniques are nothing new. This paper seeks to establish the conditions under which models were trained before the most recent EU copyright interventions, such as the exceptions in the Copyright in the Digital Single Market Directive (CDSM)[3] of 2019 and the tailored provisions in the proposed AI Act.[4]

[2] While there are legal actions claiming copyright infringement of generative AI inputs in many jurisdictions, US "fair use" jurisprudence is likely to be the global trend setter. Key live cases at the time of writing include: *Getty Images (US) v. Stability AI*, U.S. District Court for the District of Delaware, No. 1:23-cv-00135; *Andersen et al v. Stability AI, DeviantArt and Midjourney*, U.S. District Court for the Northern District of California, No. 3:23-cv-00201 (class action of visual artists); *Silverman v. OpenAI and Meta*, U.S. District Court for the Northern District of California, No. 3:23-cv-03416 (class action of writers); *Authors Guild v. OpenAI*, U.S. Southern District of New York, No. 1:23-cv-8292 (class action of trade body with a group of famous writers).

[3] Directive (EU) 790/2019 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending the Directives 96/0/EC and 2001/29/EC OJ L 130/92 2019 (CDSM Directive). Art. 3 Text and data mining for the purposes of scientific research; Art. 4 Exception or limitation for text and data mining for all purposes subject to opt-out.

[4] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts,

We adopt an inductive approach, attempting to get close to the "real world" of machine learning. Through a detailed empirical description of a selection of cases (in a social science sense), we seek to explore the legal issues that are involved.

The selection of sites for case analysis poses its own generalisability challenge. In case study research, we need to reflect on why selected empirical settings are more or less reflective of the phenomenon under investigation, i.e. rapidly evolving data analytic technologies.[5] In consultation with scientific researchers and technology companies, we identified in 2020 three case studies that together reflect a range of techniques and processes that underpin advances in machine learning.

1. Machine learning for scientific purposes, in the context of a study of regional short-term letting markets;
2. Natural Language Processing (NLP), in the context of large language models;
3. Computer vision, in the context of content moderation of images.

The selection took account of the EU policy objective of supporting innovation in this field, covering different purposes (such as scientific research or applied industrial uses) and different media modes (such as texts and images).[6]

In the study of cases in a legal context, there is a further tension between an unstructured approach that inductively offers rich descriptions from multiple sources (such as public documents, observations, interviews) and the need to capture the empirical world in a form recognisable for subsequent legal analysis. In law, this challenge of "fact-finding" is typically discussed under the concept of evidence.[7] In legal disputes, there is an assumption that a representation of facts can be settled (typically in first instance cases). It is then the application of rules to the facts that may be the subject of appeals. The case studies presented in this paper offer such a possible description of facts that will aid the development of legal analysis and policy recommendations.

The case studies were initially researched between October 2020 and July 2021, and updated in September 2023. They rely on publicly available sources (published

Footnote 4 continued
Brussels, 21.4.2021 COM(2021) 206 final; Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(1). The Parliament text includes a new obligation for providers of foundation models to "document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law" (Art. 28b). The political agreement reached in trilogue between the Commission, Parliament and Council on 9 December 2023 preserves some of this wording. See further discussion in section 4 below.

[5] For a classic account of the selection problem in case study research, *see* Seawright and Gerring (2008).

[6] *See also* Recital 8, CDSM Directive: "As research is increasingly carried out with the assistance of digital technology, there is a risk that the Union's competitive position as a research area will suffer, unless steps are taken to address the legal uncertainty concerning text and data mining."

[7] Ho (2015).

scholarly papers, official information issued by companies, policy documents, official reports) and expert feedback.[8]

3 Cases for the Study of Copyright and *Training Data* in Selected Machine Learning Environments

3.1 Machine Learning for Scientific Purposes, in the Context of a Study of Regional Short-Term Letting Markets

For machine learning projects, researchers usually start with defining the problem, choosing the data sources and algorithms, with the aim of eventually releasing a trained model (deployment stage). The stages in between can be characterised as data collection (which often involves scraping), data processing, training (supervised or unsupervised) and the output stage – such as rental predictions, language understanding or audio-visual content moderation.

The technological developments that underpin machine learning (ML) include so-called "deep learning" as a form of ML where multiple artificial neural networks carry and interpret complex raw data. With more layers, the trained model becomes more likely to solve complex problems but this also leads to less clarity about why the AI system responds, reducing the explainability and interpretability of outcomes.[9] Generative adversarial networks (GAN) comprise two deep learning networks (one generator and one discriminator), and they learn by competing with each other. GANs can be supervised and unsupervised.[10] Although neural networks were proposed as early as 1943, research on neural networks and the use of deep neural networks have increased dramatically during the last decade with the availability of cheaper computational power and resources.[11]

3.1.1 Collection Stage

For the first case study, we investigated scientific research on changes in housing markets relating to short-term rental letting services. Most property listing websites, such as AirBnB, do not create a new page for every listing. Instead, they use a template that is automatically filled with data for that specific property as entered by the users, such as a property owner, seller or host. The data available on AirBnB specifically include property descriptions, user reviews, photographs of the property (saved only as hyperlinks), location, longitude and latitude of the property,

[8] As part of the identification and validation process for the case studies, scoping discussions were held with research groups in urban studies, natural language processing and computer vision, culminating in a formal review workshop held at the University of Glasgow on 27 May 2021. The primary material is documented here: https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/. While there have been rapid technical advances in machine learning over the period since, the lifecycle methodology identified in the initial research has remained stable (CMA 2023).

[9] Birhane et al. (2023).

[10] Miller (2019), chapter 10.

[11] Seifert et al. (2017).

neighbourhood ID, available dates, maximum and minimum price, place type and number of guests and user ratings.

Scraping involves manually or automatically collecting data from websites. Screen scraping involves scraping the data that is displayed on users' screens. Web-scraping or web-harvesting involves collecting all underlying data from a website, including website scripts. Web-crawling can be defined as "accessing web content and indexing it via hyperlinks";[12] only the URL but no specific information is extracted.

There are multiple ways of categorising scraping tasks. A typical sequence includes: (i) accessing the web pages, (ii) finding specified data elements, (iii) extracting, (iv) transforming, and (v) saving these as a structured data set.[13]

If the research design and specific collection purpose is still under development, researchers often collect all available information. At this stage, no distinction may be made whether particular data is created by AirBnB or uploaded by the property hosts. Screen scraping is limited to what is available to the visitors, web-harvesting targets and collects all data, and web-crawling follows and indexes all links (those that can be visited and scraped). Since scraping relies on how data is displayed, even small changes in the display of the website can disrupt the collection stage.[14]

A standard method for streamlining the scraping process is the use of an application programming interface (API). APIs need to be made available by the service provider. The stages of using an API for data collection may be summarised as follows: (i) finding the API and exploring functionality, (ii) registering for API use and retrieving keys, (iii) calling the API to collect data, and (iv) processing the data.[15] API scraping may not result in access to previously inaccessible data, but it speeds up the process by circumventing the rendering stage.

Developing and maintaining APIs requires substantial resource investment by service providers.[16] In fact, many websites do not make their API openly available in order to prevent competitors from gathering business intelligence. Although researchers are not directly competitors, they are often unable to access efficient APIs and have to rely on different scraping strategies.

In the example of AirBnB, their API is not openly available to the general public, but may be requested by developers and certain groups of users, such as hosts wanting to use their own interface to add multiple listings at once or external partners such as travel companies and e-commerce firms (e.g. Groupon).[17]

In addition to concerns about loss of control over the data and its devaluation, website operators have to make sure that any scraping does not cause system overload. Excessive requests from the same Internet Protocol (IP) range are often blocked to ensure server stability. Since data hosts can detect unusually high or repetitive tasks from the same user account (if scraping is performed after the login

[12] Campbell (2019); Jennings and Yates (2009); Hillen (2019).

[13] Boeing and Waddell (2017).

[14] Hirschey (2014).

[15] Munzert et al. (2015).

[16] Massimino (2016).

[17] Lunden (2017).

page) or from the same IP address, researchers typically use proxies to distribute their requests to avoid exceeding this threshold and being blocked.[18]

Additionally, many websites have terms and conditions that restrict the collection and analysis of their data. Under AirBnB's Terms of Service (both for European and non-European users), there are terms that limit the ways and purposes of using the platform. For example, under section 12.1 of the Terms of Service (reviewed in September 2023), the following is not allowed: "scraping, hacking, reverse engineering, compromising or impairing the platform, using bots, crawlers, scrapers or other automated means, attempts to circumvent any security or technological measure, taking any action that could damage or adversely affect the performance or proper functioning of the platform". Furthermore, the content cannot be used without the permission of the content owner and can only be used as necessary to enable the website to be used as a guest or host.[19]

3.1.2 Processing Stage

After the targeted data is collected, it is then structured in a manner that is suitable for the identified research purposes. With the increase in computational power and reduction in storage costs, it has been suggested that researchers are now able to scrape more data and can choose to be less conservative. This also implies that more data are to be filtered and cleaned.[20]

As the property information in this case study is added by the users, it can be messy, and the researchers might have to go through substantial wrangling and validation to make the data usable. For example it will be necessary to identify and remove duplicate listings (by relying on property ID, location and the size of the property) or identifying mistakes such as typos in the rental price.[21] As part of data validation, researchers have to ensure that the collected data is reliable and usable for their purposes.

It is also possible to enrich scraped data with data from other sources. For example, there are websites and analytics companies based in the United States, such as AirDNA and SmartHost, that collect and aggregate AirBnB data to guide the hosts and nearby businesses. There are also US sources that provide scraped data together with their own analysis. Researchers, both inside and outside the United States, often rely on such scraped datasets, commentary and research outputs by such third parties.[22]

3.1.3 Analysis and Output Stage

The collected data can be one-off and reflect a particular point in time or it can allow real-time updates (such as price comparison websites).[23] There is a growing body of academic literature based on AirBnB. A wide range of issues are addressed,

[18] Hirschey (2014).

[19] AirBnB (2023).

[20] Gold and Latonero (2018).

[21] Boeing and Waddell (2017).

[22] Scassa (2019).

[23] Hillen (2019).

such as the extent to which neighbourhoods are vulnerable to the switch from long-term letting to short-term letting.

Examples of papers applying machine learning techniques to scraped AirBnB data include studies of short-term rental markets in Corsica[24] and New York.[25] The former investigates the pricing of short-term vacation rentals on the island of Corsica for the years 2016 to 2019 using data from the US headquartered commercial service AirDNA (with European services based in Barcelona), which appears to have a commercial relationship with AirBnB and access to its API.[26] The latter uses data scraped by InsideAirBnB, a public interest project to improve housing policy that appears to rely on US law to assemble AirBnB data without permission.[27]

The results of the analysis are shared in formats chosen by the researcher (such as journal articles, reports, heat maps). The extent of the data used in these publications varies case by case. Some outputs may be complementary to AirBnB services, encouraging use of its services. Others may provide uncomfortable evidence that may convince policymakers to restrict AirBnB properties in certain cities or regions.

3.2 Natural Language Processing (NLP), in the Context of Large Language Models

Natural language processing (NLP) is located at the intersection of computer science and linguistics. It is a form of machine learning where the purposes can range from analysing larger texts to computers generating realistic texts. The applications of NLP include information extraction, machine translation, sentiment analysis and, most prominently, natural language generation via powerful large language models such as OpenAI's GPT.[28]

NLP can be supervised or unsupervised. Supervised learning requires labelled/tagged text data, with an "annotation" stage in their workflow. Unsupervised NLP uses unlabelled data and instead detects patterns, but it requires very large datasets. If some labels are generated by humans and others are not, then the process will be classified as semi-supervised machine learning – which is useful for projects holding small annotated datasets together with large amounts of raw data found online.

[24] Brunstein et al. (2023).

[25] Kalehbasti et al. (2021).

[26] AirDNA explains its policy on "scraped data at scale" as follows (https://www.airdna.co/airdna-data-how-it-works): "From daily calendar pricing to cancellation policies and booking lead time, we aggregate and process comprehensive data on over 10 million properties in over 120,000 international markets. We do this by 'scraping' (or extracting) data using a host of servers. The process is 100% legal, and our relationships with these booking platforms are strong and mutual."

[27] http://insideairbnb.com/.

[28] ChatGPT, a general purpose chatbot built on a large language model developed by OpenAI, was released to the public in November 2022. Microsoft is the largest investor (with an exclusive licence to the technology). In February 2023 OpenAI's GPT-4 model was integrated into Microsoft's search engine Bing.

NLP research focuses on achieving and improving various tasks. Some tasks have direct applications, such as translation or summarisation. Other tasks such as segmentation or named entity recognition are used to inform other tasks and turn the texts into machine-readable data.

3.2.1 Collection Stage

The first step for NLP is the compilation of the necessary data. The data can come from anywhere, ranging from user comments to ancient philosophy. The data collection stage is similar to the scraping process described in case study one: the necessary data is identified in line with the research purpose and then it is targeted with the appropriate data collection methods.

A prevalent source of training data are freely available online materials, such as the books from Project Gutenberg or the Spoken Wikipedia.[29] NLP researchers may also choose to focus on licensed corpora[30] or scholarly literature held in databases to which they have access.[31] Large language models seem to rely on the collection of the whole of the public internet.[32]

3.2.2 Pre-processing

The data then goes through pre-processing. This part involves different tasks in order to understand the texts. The collected material goes through some changes at this stage, which will be important for the legal analysis later. First, common formats such as PDF or Microsoft Word need to be converted into text for the NLP tasks that follow.[33]

Tokenisation separates texts into smaller units in a way that can be read by the machine. These smaller units can be word pieces or characters. Parts of speech (POS) tagging is when words are tagged as noun, verb, or prepositions. Normalisation removes variations that are not important for the final research target. Normalisation includes tasks such as lemmatisation, stemming or spelling correction, which all change the text. Stemming removes the end of the word, while lemmatisation changes the word into its base or dictionary form. Such tasks are sometimes performed by an algorithm, but humans can be consulted as well, at least while these methods are being developed or applied to new application domains.[34]

3.2.3 Training

The stages after pre-processing then differ according to the type of the learning.

[29] https://www.gutenberg.org/; https://en.wikipedia.org/wiki/Wikipedia:Spoken_articles.

[30] Eckart de Castilho et al. (2018).

[31] Przybyła et al. (2016).

[32] Reisner (2023); Schaul et al. (2023).

[33] Cottman (2020).

[34] Jurafsky and Martin (2020).

(a) *Supervised*: If the project relies on supervised learning, pre-processed data is annotated by humans. Data that was previously unreadable to the machine becomes usable through the annotation stage. During the annotation process, it is possible to both add annotations to the original text or create a separate file for annotations.[35] The former has the advantage of keeping both the text and annotations in a single file – such as an XML file – so the NLP algorithms have access to both.

(b) *Unsupervised*: If unsupervised, learning requires no human input once the data is collected. There is no annotation stage. The project could involve multiple tasks that support each other by creating annotations, but as long as NLP relies only on pre-trained models and the final task does not involve humans, it would still be characterised as unsupervised training.

Although unsupervised learning is possible and is a growing field in NLP, it is not widely accessible to smaller groups due to large data requirements and the need for computer power. Companies that have such resources, such as Google or OpenAI, use it to create pre-trained models.

Pre-trained embeddings and models are trained on a large corpus in an unsupervised manner, then fine-tuned in a supervised manner.[36] These are then made available for other users, so that they can be used to support other supervised and semi-supervised learning projects, skipping some stages in collection and pre-processing. This may result in a small number of dominant players in language modelling.[37]

The following paragraphs will explain where embeddings and models sit within the developments of NLP. It is useful to take such developments into consideration for our legal analysis, as the approaches determine the amount and type of data that is used and the parties' involvement.

– In earlier NLP projects, a "bag of words" approach assigns a unique token to words, in order for a text to be displayed in numbers. For the transformation of words to numerical representations (vectors), the basic method is to count how many times a word occurs in a text, without paying attention to the order of the words. Since this approach would identify words such as "the" or "is" as the most common and therefore the most important, the weighting of the words needs a separate adjustment (TF-IDF encoding). N-grams extract a consecutive n-number of words from the text for analysis.[38] These methods are still used, but are now supported by the others below.
– Word embeddings (2013 onwards): Embedding models mean giving vectors that show the connection between words. This allows the machines to understand which words go together, which helps in tasks like prediction or translation. There are word embedding models such as *word2vec* (by Google) and *GloVe* (by Stanford).

[35] Przybyła et al. (2016).

[36] OpenAI (2018).

[37] Soper (2020).

[38] Tan (2020).

The researchers then have the option of either (i) relying on pre-trained word embeddings (based on the training done by their developers) such as *word2vec* trained on the Google News corpus,[39] or (ii) training the embeddings themselves to make sure that they assign numerical values based on their specific dataset/research topic – so that they can be used on later NLP tasks with greater accuracy.

Since the first option is trained on generic texts, they are not overly helpful for use on very specialist texts, for example legal documents.[40] This means that researchers of specific topics still might prefer to train their own word-embedding models with their own training data. The fact that pre-trained embeddings rely on easily found text material also leads to bias problems. For example, *word2vec* carries the same gender biases present in the news corpora it was trained on.[41] But since researchers can only view the trained *word2vec*, and not the news corpus it was trained on, it is also hard to pinpoint the reasons for this bias or to adjust outputs.[42]

Language models (2018 onwards): The most recent models rely on deep learning. They also excel in analysing the whole document, but here the vectors are dynamic and adapt to the context. Transformer models are able to understand the difference when the same word is used in different contexts.[43]

– Large language models rely on deep neural networks, which are better at detecting and predicting "complicated linguistic structures along with their long-distance relationships, as humans do".[44] Another difference of transformers is that they can process words "in parallel", instead of "sequentially one by one" like the former methods. This increases the speed in processing large amounts of data.

Transformer models are trained on unlabelled data, for example Google's BERT trained on the English language Wikipedia and the Brown Corpus.[45] They can then be tweaked for other tasks. One of the drawbacks is that they do not exist for all languages. Additionally, the pre-trained versions might still require some fine-tuning. They might not be sufficient on their own, but they can make smaller projects viable.

3.2.4 Trained Model

The final stage is the creation of the trained model (a permanent file). Once the researchers have a trained model, they can use it on previously unseen datasets or use it to inform and support other larger tasks. What the trained model achieves

[39] https://code.google.com/archive/p/word2vec/

[40] Chalkidis and Kampas (2019), pp. 171, 174.

[41] Buonocore (2019).

[42] Levendowski (2017).

[43] Vaswani et al. (2017).

[44] Chalkidis and Kampas (2019).

[45] Google (2019).

depends on what task it was trained for. Some tasks have direct applications, while the others mainly help other NLP tasks.

Algorithms developed for Natural Language Understanding aim to determine the meaning of a sentence. AI applications use syntactic and semantic analysis to "read" the text. Document classification, sentiment analysis or named entity recognition are examples of such "understanding" tasks. Algorithms that "write" or "speak" are labelled Natural Language Generation, or in popular parlance generative AI.[46] For example, machine translations or chat bots that answer questions achieve both understanding and generation through multiple NLP tasks.

It is not possible to remove some of the data after the model is trained. If a small part of the data needs to be removed (due to copyright or another reason, for example following an injunction), then the whole model may need to be retrained from the beginning or the output "aligned".[47]

3.3 Computer Vision in the Context of Content Moderation of Images

The third case study focuses on computer vision. The developments in this field have been largely driven by industry uses, such as facial recognition or self-driving cars.[48] The discussion here will focus on the use of object recognition technology for content moderation.

In supervised learning, models are trained with annotated datasets, and also receive human feedback when wrong classifications are made based on the features presented. In unsupervised learning, algorithms learn by looking at different images and recognising similarities, as humans do by observation.[49] The earliest use of deep neural networks was in the field of computer vision.[50] An example of applying deep learning is the use of generative adversarial networks (GAN) in creating art. In this unsupervised form of learning, the generator continuously tests the discriminator with realistic works. In addition to requiring large datasets of images,[51] such practices lead to questions about the copyright status of AI-created works (which is outside the scope of this paper).

Although computer vision tasks vary widely, the process starts as in the previous case studies with the collection of input data, followed by the processing of the data (which are different from NLP pre-processing tasks), followed by training and deployment to outputs (which could range from a simple yes/no classification decision to a detailed, machine-generated response).

[46] Kavlakoglu (2020).

[47] Zhang et al. (2023).

[48] Arnold et al. (2019).

[49] Miller (2019).

[50] Seifert et al. (2017).

[51] 75753 paintings were used to train the Generative Adversarial Network in the project where creative adversarial networks were proposed for the first time. Elgammal et al. (2017).

3.3.1 Data Collection

The images or videos can come from various sources, such as phone cameras or medical devices. When training a computer vision model, it is important to use a dataset that is similar to the data it will be used for. For common objects, there are open datasets of labelled images online. One of the earlier projects of computer vision, ImageNet, was launched in 2007 and holds over 14 million images labelled by participants. LAION's 2021 open dataset consists of 400 million image text pairs (in English).[52] Easily accessible datasets are not sufficient for very specific research problems nor do they give any competitive edge if everyone trains their AI systems with the same images. Another option is using own image data or even a digitally generated dataset (synthetic data). If the collected data is too small, it can be augmented.

3.3.2 Pre-processing

Once the data is collected, the images or videos go through pre-processing tasks, which are relevant for the legal analysis. One of the tasks in pre-processing is the resizing of the image, so that all images in the dataset are the same size. Converting colour images to grayscale reduces the computation complexity, for research problems where the colour does not matter.[53] Another task is noise reduction where the background features are smoothed and removed, so that the machine can focus on a single feature.[54]

It is possible to increase the dataset and prepare an AI application for recognising the same objects in different environments by data augmentation. This can be achieved by rotating, scaling, cropping or flipping the image. While augmentation follows similar steps as above, it is only applied to the training data sets and not to the test sets.

3.3.3 Training Stage

Similar to NLP, Computer Vision has supervised, semi-supervised and unsupervised training options. Supervised and semi-supervised require annotated datasets. In unsupervised learning, computer vision is able to recognise common features in images (cluster analysis), without annotations.

Annotation is performed by assigning a label to the selected part of the image, or a single label for the entire image. Feature extraction can be included under this stage – or alternatively be seen as a separate stage in the computer vision process. A feature is defined as "a measurable piece of data in your image that is unique to that specific object … a distinct colour or a specific shape such as a line, edge, or image segment".[55] The features can be extracted manually or automatically. The training then occurs based on the extracted features.

[52] http://www.image-net.org/about; https://laion.ai/blog/laion-400-open-dataset/

[53] Elgendy (2020).

[54] Kumar and Hosurmath (2019).

[55] Elgendy (2020).

Some steps here can be merged due to technological developments in deep learning. Convolutional neural networks (CNN) are used for image classification and recognition problems. Prior to CNNs, the standard ML training process (for videos) included (i) extracting features, (ii) combining the features into a fixed-sized video level description, and (iii) a classifier trained on "bag-of-words" level descriptions. CNNs combine all these stages.[56]

CNNs have layers of "small computational units that process visual information hierarchically in a feed-forward manner", so each layer works as an image filter and extracts a feature from the image and the image becomes increasingly more explicit along this hierarchy.[57] The process is slightly different for videos. When used for a video, AI technology has to detect key images which are the most relevant images in the video and eliminate redundant or blurry images. This simplifies the subsequent analysis work.[58] CNNs can be used both supervised and unsupervised, and although widely used for image classification, they can also be used for text classification.[59]

3.3.4 Models for Content Moderation

Trained models can be used in tasks such as image classification (used for example in medical diagnosis or reading traffic signs), object detection and localisation, generating images, face recognition and image recommendation.[60] Some computer tasks vision are more suitable for unsupervised methods (such as image classification), while others might require more human input.

When using AI for content moderation, it is possible to combine computer and human moderation: for example, when determining if user generated content is harmful. An AI application can flag content as "uncertain", which then goes to human moderators whose decisions can be fed back as training data for the AI to learn how to address similar images or videos.[61] Trained on datasets for recognising for example nudity, violence or drugs, machine learning technology is being used by various companies for content moderation.[62]

It should be noted that in using computer vision for content moderation, machine learning is only one of the methods. Other methods include hashing and fingerprinting. Hashing works by generating unique identifiers for files and then comparing these with reference databases for detecting e.g. terrorist content or viruses. Fingerprinting is similar to hashing, with the unique identifier not based on the file but on characteristics of the content.[63] While it is easier to match content found online to previously flagged content, training models to make decisions on new content is more difficult. Furthermore, the reasoning behind machine learning

[56] Karpathy et al. (2014); Cambridge Consultants (2019).

[57] Gatys et al. (2016).

[58] CSPLA, CNC and HADOPI (2020).

[59] Guérin et al. (2018).

[60] Elgendy (2020).

[61] Sartor and Loreggia (2020).

[62] Examples include Clarifi, Amazon Rekognition, Valossa and Sightengine. *See* EUIPO (2020).

[63] EUIPO (2020) pp. 7, 15.

decisions is more obscure.[64] At this stage, AI technology is mainly used for improving and making fingerprinting faster. It is not sufficient on its own for e.g. copyright content moderation.[65]

4 Legal Analysis

The three case studies identify the sourcing and processing of input data as the critical first element in the machine learning lifecycle. The nature of training data requires an assessment of the legal rules governing such data. In this section we attempt to clarify the legal ambiguity of the term *data* in "training data", focusing on copyright and related rights.

4.1 Copyright, Uncertainty and AI Development

As shown in the case studies, content such as texts and images are common training data. However, when expressed in an original form, they also become natural candidates for copyright protection as literary or artistic works. Copyright theory traditionally distinguishes protected works from unprotected material. The former are original expressions in the literary and artistic fields. The latter is a broad category which does not warrant protection for various reasons: lack of originality, lack of (a stable and objective) expression, expiry of the term of protection or other more specific reasons for which we refer to our previous analysis.[66] Alongside copyright, there are other rights that protect activities that are related to the creative process but that do not accrue to the level of works of authorship. Examples are phonograms, broadcasts, performances and, particularly relevant for present purposes, the EU Sui Generis Database Right (SGDR).[67] This is a special form of protection for databases against acts of extraction and reutilisation of substantial amounts when the obtaining, verification and presentation (but not the creation) of the database required a substantial investment.

As a first approximation it can be stated that most of the literary and artistic works found on the internet, at least those created in the last 70 years, are protected by copyright. For related rights the term of protection may vary. For the SGDR it is 15 years, which can however be renewed potentially indefinitely as long as there have been new substantial investments.[68] These are often, albeit not always, the same resources that are used for text and data mining as well as for more recent advancements like "generative AI".[69] As seen in the first case scenario, web

[64] CSPLA, CNC and HADOPI (2020).

[65] EUIPO (2020); Margoni et al. (2022).

[66] Margoni and Kretschmer (2022), p. 688.

[67] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, Official Journal of the European Communities, L 77/20.

[68] Art. 10 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive); *See* also Hugenholtz (2016), Art. 10, notes 3 and 4.

[69] Emanuilov and Margoni (2023).

crawlers are commonly used to analyse and archive web resources which are then distilled into custom-made datasets or corpora to be fed to learning algorithms. Case studies two and three show in detail how the collected data is processed and turned into a trained model which will form the knowledge basis for the AI application to process the requested query by a user and deliver the output in the form of a translation, a text completion or a more complex literary or audiovisual work in the case of generative AI.

The legality of these practices has often been assumed, mainly relying on fair use principles, both within the US as well as in other jurisdictions. However, at closer look, the law appears far less clear than research and industry practice may suggest.[70] Legal uncertainty represents fertile ground for borderline practices to emerge, such as where commercial AI developers are told by their legal departments to "mine everything and then destroy the training material" since it will be very difficult to reverse-engineer the trained model, go back to the training material and prove infringement.[71]

These practices take advantage of the underlying legal uncertainties and the ensuing unregulated power imbalances to extract, accumulate and concentrate value from data. A striking example of this effect is the emergence of a handful of so-called "foundation" models[72] that are developed by the few large tech corporations which have access to the necessary data and can afford the uncertainties and costs of potential copyright litigation.[73] Such short-term accumulative practice enabled by legal uncertainty and performed by vertically integrated firms may consolidate a techno-economic oligopoly. It has the additional effect of delaying an evaluation of the long-term legal, economic, social, cultural and environmental sustainability of what has been described as a form of data extractivism.[74] The EU has taken a pioneering stand in this area by proposing a set of novel regulatory solutions.

4.2 The Role of Copyright in the AI Lifecycle

Foundation models, including the popular large language models (LLMs) such as OpenAI's GPT3&4, Google's PaLM, or Amazon's Alexa TM, as well as text-to-image models such as Midjourney or Stable Diffusion, are trained on a wide array of publicly available materials which are probably protected by copyright. When this is the case, acts of training often require authorisation under EU copyright law. The reason is to be found in the broad definition and interpretation of the right of reproduction.[75] In other words, given the many copies needed to perform acts of

[70] For a comprehensive analysis of the law of data scraping in the UK, conducted in preparation for the current study, *see* Burrow (2021).

[71] Project review workshop of 27 May 2021. The primary material is documented here: https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/.

[72] Bommasani et al. (2022), p. 4.

[73] Margoni et al. (2022).

[74] Couldry and Mejias (2019).

[75] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2002] OJ L 167/10, Art. 2 ("InfoSoc Directive"); Case C-5/08 *Infopaq International A/S v. Danske Dagblades*

training, and given the fact that the EU law broad definition of reproduction arguably covers most of those copies, then acts of training (i.e. of copying) need authorisation even when they are mere temporary and incidental copies.

Authorisations may take various forms but usually they possess either a statutory (exceptions and limitations) or a contractual (licences, individual, public or collective) nature, and sometimes a mix thereof (e.g. statutory licences). Starting with the statutory forms of authorisation, it can be observed that within the EU framework, there are several potentially relevant exceptions. Of particular relevance for present purposes are Art. 5(1) of the InfoSoc Directive (ISD) and Arts. 3 and 4 of the Copyright in the Digital Single Market Directive (CDSM).[76] As more extensively discussed elsewhere,[77] the temporary copying exception of Art. 5(1) has historically represented the balancing mechanism between the protection of rightholders' interest on the one hand and the right of users to technological development and innovation on the other hand. This is visible both in the legislative history of the provision[78] as well as in the more recent interpretation offered by the CJEU.[79] Article 5(1) however is limited in various ways, chiefly in that it is an exception only to temporary acts of reproduction, thus permanent copies – which are fundamental for the replicability of machine learning results – are excluded from its scope. Other conditions of Art. 5(1), such as that of lawful use, contribute to reducing the suitability of this provision for modern text and data mining (TDM) processes even within temporary reproductions.[80] Its role, however, should not be completely disregarded. The fact that the CJEU has confirmed that it applies to cases of (commercial) information extraction and retrieval services may suggest renewed relevance in the context of the opt-out in Art. 4 CDSM.

4.3 Opt-Outs and Temporality

Regarding Arts. 3 and 4 CDSM, we refer to our previous study.[81] It is important to note however that the empirical cases suggest a differentiated categorisation of the lawful access role in the opt-out processes.[82] As usually reported in the literature

Footnote 75 continued
Forening [2009] ECR I-6569, ECLI:EU:C:2009:465, paras. 42, 43 and 47; Margoni and Kretschmer (2022), p. 690; Joined Cases C-403/08 and C-429/08 *Football Association Premier League v. QC Leisure and Karen Murphy v. Media Protection Services* [2011] ECR I-10909, ECLI:EU:C:2011:631, para.159. *See* also Geiger et al. (2018a) and (2018b); Ducato and Strowel (2019); Otero (2021); Rosati (2018); Guadamuz and Cabell (2014).

[76] *Ibid.* and footnote 3 above.

[77] Margoni and Kretschmer (2022), p. 690.

[78] InfoSoc Directive, Recital 31: "A fair balance of rights and interests between the different categories of rightholders, as well as between the different categories of rightholders and users of protected subject-matter must be safeguarded. The existing exceptions and limitations to the rights as set out by the Member States have to be reassessed in the light of the new electronic environment."

[79] Case C-5/08 *Infopaq International A/S v. Danske Dagblades Forening* [2009] ECR I-6569, ECLI:EU:C:2009:465, paras. 56, 57 and 59.

[80] Margoni and Kretschmer (2022), p. 693.

[81] *Ibid.*

[82] For a detailed analysis of the relationship between lawful use, lawful user and lawful access *see* Synodinou (2019).

(including by the present authors), one of the main differences between Art. 3 and 4 lies in the imperative nature of Art. 3, i.e. it cannot be limited by contract. Instead, Art. 4 operates as an exception only if rightholders have not reserved the right to TDM in the form prescribed by the law.[83]

It is arguable that in specific sectors characterised by a strong concentration of the supply side (for instance the short-term rental market services of case study one, but also other fields such as the commercial scientific publishing industry), the requirement of lawful access may very well operate as a form of (surreptitious) reservation of the right to TDM. In other words, if the supply side is sufficiently concentrated, there is an inelastic effect on the demand. Researchers cannot operate without access to the knowledge found behind the paywalls of vertically integrated platforms, such as those operating rental or publishing services. Rightholders are under no obligation to make that wealth of data accessible. They can decide whether to do so and under what conditions. If they do, however, they cannot limit – or in economic terms – segment that offer. Access implies TDM. No access implies no TDM. Under these conditions, the real effect of Art. 3 is simply to rule out a third option: access without TDM (or for an additional price).

As emerged from the analysis of case study one, services often allow access to their datasets via their Application Programming Interfaces (APIs). Whereas in the most traditional sense APIs establish the standards for two computers to communicate, their design often embeds choices that determine access conditions. These conditions may be of different nature and often include limitations necessary for the security and stability of the network or databases (as allowed by Art. 3(3) CDSM). At other times, however, APIs may limit quite substantially what users can do. In other words, it is technically rather simple to design an API that only allows a certain number of requests, or certain lengths or complexity of queries, or again a certain search and retrieve function. It is difficult to state when these limitations will pass that red line between security measures allowed by Art. 3(3) and become a form of (forbidden) limitation of the rights established by Art. 3. It is clear that this techno-legal uncertainty, combined with the power asymmetry characteristic of certain markets may *de facto* operate as a form of circumvention of the imperative nature of Art. 3. In practice, business models are emerging where alongside a basic access (with TDM) via APIs, there is a "premium" access (with TDM) via APIs that allow more freedom in setting the search and analysis parameters.[84]

This reconstruction should not be entirely surprising. The impact assessment of the CDSM had identified the role of "lawful access" as a condition allowing commercial scientific publishers to retain their licensing business models. However, accepting this effect leads to the necessary conclusion that the difference between Arts. 3 and 4 in terms of opting out are more temporal (when), rather than existential (whether). TDM for scientific purposes can be limited. The main difference from Art. 4 is that this form of TDM is bundled with access. Rightholders make the

[83] Art. 4(3) CDMS provides that the exception applies on condition that use of works or other subject matter "has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means".

[84] Schirru and Margoni (2023).

decision to allow TDM in the context of their decision to allow access to their databases. Access to the databases can be subject to a number of monetary and non-monetary conditions. The only condition that cannot be enforced is to forbid or charge extra for TDM. However, as seen, even this prohibition can be circumvented, at least partially, via a techno-regulatory (ab-)use of APIs.

When the opt-out from TDM is performed, either simultaneously with the decision not to grant access under Art. 3, or successively in the case of Art. 4, the next question is, usually, how to monetise it. Licences are a common answer to the question. Contractual models specifically geared to the licensing of "TDM" or "AI" uses are likewise emerging in practice.[85] Before moving to a brief overview of the role of licences, however, it is important to note that the formalistic interpretation adopted in EU law that classifies copies in machine learning as a form of copyright relevant reproduction is not necessarily embraced outside the EU or in copyright theory. Concepts such as "non-consumptive uses" proposed in the scholarship may find a fertile ground in legal systems that either follow a utilitarian view of copyright (e.g. US, Canada, Singapore),[86] or that have identified computational uses as a key policy priority for domestic technological development (e.g. Japan).[87]

4.4 Licences and the AI Lifecycle

Regarding the contractual forms of authorisation, various scenarios may be envisaged: direct licences, either individually negotiated or publicly offered as standard public licences; collective licences, mandatory licences or even forms of fair compensation. Regarding direct licences, there appears to be renewed interest in the possibility for authors to individually negotiate a "right to train" with AI developers, also thanks to the opt-out provisions of Art. 4 of the CDSM Directive. A *TDM.txt* or *AI.txt* file, replicating in this new environment the workings of the more traditional *Robot.txt,* have been proposed.[88]

The ambition to charge a substantial fee for a single work however seems difficult to achieve, since large models are commonly trained on billions of words.[89] While collective management seems to be a possible avenue, there is currently no working model that could offer an economically efficient infrastructure for such micro-uses and payments. As an alternative to the (problematic) practice of data scraping or accessing openly licensed data sets, commercial publishers or commercial stock image services offer "AI training licences" not to individual works, but to their entire databases containing hundreds of thousands of works.[90] An

[85] *Ibid.*

[86] Flynn et al. (2020); Sag (2019); Craig (2017).

[87] Ueno (2021).

[88] There are artist led initiatives, such as https://spawning.ai/, that selectively restrict or permit the use of online content for commercial AI training. Google, Microsoft and OpenAi have all developed their own proprietary opt-out protocols. *See* Keller (2023) for a critique of model specific opt-out mechanisms.

[89] ChatGPT-4 was trained on 570GB of data and 300 billion words. *See* Hughes (2023).

[90] Schirru and Margoni (2023).

alternative and interesting option has been recently proposed in the literature and focuses on a type of flat fee applied to AI firms, which would then be redistributed to rightholders.[91]

There are also circumstances where contracts acquire a different, more pervasive role. Situations where the underlying material is not covered by copyright or related rights are conceivable. In these situations, contracts perform a different function. They do not simply represent the authorisation to perform an act that would otherwise be reserved by copyright law. Characterised by the absence of an underlying property right, contracts may very well set the boundaries of what is allowed and what is not, in ways that can go even beyond the default under copyright. In fact, whereas copyright has the advantage of offering an *erga omnes* underlying right to which the contract becomes the only use-enabler – thus somehow adding a sort of limited third-party effect to contracts – it also embeds a balancing of interests (e.g. exceptions and limitations) that in certain cases have an imperative nature that cannot be limited by contract. This does not happen often (and it is ultimately a matter of domestic law in the EU), but there are cases where it is clearly stated that a certain exemption cannot be overridden by contract. Examples are found in the Software Directive, in the Database Directive and, importantly for present purposes, in the CDSM Directive with regard to Art. 3. However, when there is no underlying property right, the contract (if enforceable) can regulate the performance between the parties in a way that the law would not have allowed had copyright existed. This interpretation was accepted by the CJEU, at least in relation to databases, in the Ryanair case,[92] where the absence of an underlying *sui generis* database right (SGDR) led the court to confirm the enforceability of terms of use that would not have been acceptable had an SGDR existed.

4.5 AI Regulation in the AI Lifecycle

The EU legislator is negotiating the challenging field of technology governance via a mix of regulatory approaches. Alongside the more familiar field of copyright law, another emerging approach is found in so-called "data and digital legislation". Examples in this field are initiatives such as the Data Governance Act (DGA)[93], Data Act (DA)[94], Digital Services Act (DSA)[95] and most relevant for present purposes, the AI Act Proposal.[96]

[91] Senftleben (2023), p. 3.

[92] Case C-30/14, ECLI:EU:C:2015:10 of 15 January 2015 (*Ryanair*).

[93] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act).

[94] Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair assess to and use of data COM/2022/68 final (Data Act).

[95] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance)

[96] European Commission, Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206;

A detailed analysis of the AI Act (AIA) in relation to copyright would be beyond the scope of this article. However, a closer look at some of the elements of the Proposal will offer an insight on the perceived role of copyright as a regulatory lever for machine learning. At the time of writing, the AI Act had reached the "trilogue" stage, with three texts available (Commission, Council and Parliament). Following the closed-door process of the trilogue, a political agreement was reached on 9 December 2023 and adoption is expected before the end of the current parliamentary period in early 2024.

It is important to note that the regulatory role attributed to copyright in the latest text was absent in the original proposal of the AI Act (European Commission text of 2021[97] and in the following Council text of 2021[98]). It emerged in the European Parliament text of 2023[99] as a response to so-called "generative AI". Generative AI within the Parliament text is a sub-type of so called "foundation" models, a new category in its own right. Specific to generative models is a new obligation to "document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law" (Art. 28(b)(4)(c). If enacted, these provisions will affect the legal analysis of our empirical settings, in particular case studies two (natural language processing) and three (computer vision), introducing an interesting provision that on the one hand seems to offer a way to operationalise the opt-out faculty of rightholders on the basis of Art. 4 CDSM, while on the other introducing a specific element of transparency into AI training.

5 Conclusion

We set out to investigate under what circumstances machine learning technology, in three empirical lifecycle settings, may come into conflict with, or may be shaped by, (EU) copyright law.

The case studies offered a detailed picture of what may be copyright-relevant reproductions in the context of machine learning. An important finding from all cases is the sophisticated sourcing and processing of input data required in the

Footnote 96 continued
Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(1).

[97] Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(1)

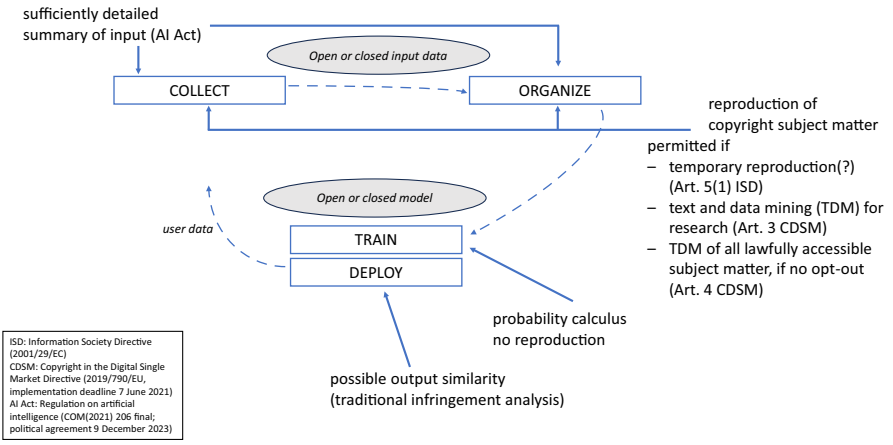
[98] Council of the European Union, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - General approach, Council Document 15698/22.

[99] Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(1)

machine learning process. The legal analysis identified deep uncertainties regarding freedoms to operate and rightholder authorisations required. There are unpredicted behavioural consequences that arise from these uncertainties, such as incentives to destroy training data. We also characterised the lawful access requirement as paradoxical, subverting the innovative aims of the text and data mining exceptions in the CDSM Directive.

The following diagram illustrates the copyright relevant stages of the lifecycle of machine learning under EU law:

Copyright representation of machine learning lifecycle (EU law)



A distinction emerges between continuously deployed models (for example, relying on current data updates) and the use of "off-the-shelf" models that may be fine-tuned or aligned for particular purposes but where data collection is essentially complete.[100]

From the legal analysis, it is clear that the right of reproduction contained in Art. 2 of the Information Society Directive (ISD) together with the temporary exception of Art. 5(1) ISD has been tasked by the Court of Justice of the European Union with the role of enabling technological development. However, we show that there is a tension in the relationship between Arts. 2/5(1) ISD and the text and data mining exceptions introduced with Arts. 3 and 4 of the Copyright in the Digital Single Market Directive (CDSM). Research use under Art. 3 is subject to the condition of lawful access (and thus contracts). The opt-out available to rightholders under Art. 4 CDSM for non-scientific purposes is a complex basis for entering licensing agreements (or for some AI firms to avoid licensing).

Predicted effects in the EU market may be summarised as follows:

– Scientific research uses, exemplified by case study one, are likely to be affected by the lack of clarity whether copying in machine learning contexts is permitted, and under what conditions. The terms of lawful access will control what

[100] Kretschmer et al. (2023).

research is possible and at what cost. Research therefore is likely to be conducted under licensing arrangements where providers of valuable data sets will set the terms for research. For example, while research or heritage organisations may have current and lawful access to a broadcasting or newspaper digital archive, rightholders may want to license that material to major AI firms for machine learning purposes and threaten to withdraw archives from settings where they may be used for public interest research. In the example of live online services, such as data about rental markets, the line between legitimate competitive control via terms of service and the public interest has not been successfully drawn with the EU's text and data mining (TDM) exceptions and the *sui generis* database right (SGDR). Here, research will likely take shelter in jurisdictions with a more permissive copyright environment, as we have seen in the case study of short-term lets.[101]

– For natural language processing (NPL) and computer vision models, case studies two and three explain in detail how information is extracted from large volumes of copyright works. Since applications of the resulting models are driven by commercial opportunities, unlicensed processing in the EU copyright framework is likely to conflict with the opt-out of Art. 4 CDSM (if use of works has been "expressly reserved by their rightholders in an appropriate manner, such as machine-readable means"). It is difficult to apply this notion retrospectively, nor may it be possible to establish the corpora of works on which specific models in circulation were trained. For future development, however, it is likely that preferential access to high quality, curated corpora of copyright works will form the basis for licensing arrangements between rightholders and AI firms.

Where does this diagnosis leave individual creators? Neither of the predicted market responses will be beneficial. Withdrawing from machine learning contexts should be possible for rightholders under the opt-out of Art. 4 CDSM, but this may reduce the diversity and quality of AI models. If licences become available, the individual creator's share of the revenues generated is likely to be minimal, since the foundation models of greatest commercial value possess billions of parameters trained on trillions of tokens (in the case of language models).[102] Creators in effect seem to demand that societies license the total sum of available human expression, for a second time. Monetary awards under this approach may be largely symbolic.

It is interesting to compare the current policy environment with the invention of the temporary copying exception to enable browsing and search during the 1990s. A broad interpretation of the exclusive right of reproduction would have undermined the viability of the Internet as a mass medium. The international legal framework was adapted to legitimise copying in web search and browsing, after the event, and many national legislators provided a temporary copying exception.[103]

[101] There are empirical indications that enhancing the compliance costs of text and data mining drives AI development towards legal systems with more permissive rules. Handke et al. (2021).

[102] Schaul et al. (2023).

[103] Agreed Statements concerning the WIPO Copyright Treaty (1996), adopted by the Diplomatic Conference on December 20, 1996.

Are there any policy options that would address our rather bleak predictions about the copyright status of input data, and perhaps move the debate to a new international consensus? We currently see three types of interventions on the table: (1) obligations to disclose copyright-relevant training data; (2) a form of collective licensing of copyright works for the purposes of machine learning; (3) legal privileges for open source models.

1. Obligations to disclose training data

In the European Parliament amendments to the proposed AI Act of 14 June 2023, a new Art. 28b, entitled "Obligations of the provider of a foundation model" provides certain additional obligations for "Providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ('generative AI')". This includes an obligation to "document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law" (Art. 28(b)(4)(c)).[104] Assuming that this "sufficiently detailed summary" will include the details (e.g. author, title, URL) of all copyright protected training material, Art. 28(b)(4)(c) aims to operationalise, at least within the subcategory of generative foundation models, the possibility for rightholders to monetise the use of their works, after they have opted out from training (or denied access) under Art. 4 CDSM. There are numerous technical issues with this proposed provision, which have been discussed elsewhere.[105] However, in our context, it is a sign of an accelerating trend towards a licensed environment we have identified above.

2. Collective licences for machine learning

Mandatory collective management has the potential to remove the risks of potential market entry and related innovation hold-ups. However, setting up a body that assembles sufficient rights across the key modi of machine learning (text, images, sound, audio-visual) may not be feasible.

Martin Senftleben suggests instead a levy approach with a focus on equitable remuneration to authors. Using the EU's Rental Directive as a model, such a levy would be paid by providers of generative AI systems to the "social and cultural funds of collective management organisations for the purpose of fostering and supporting human literary and artistic work".[106] This approach is attractive but would be bureaucratically challenging, with key issues around levy efficiency remaining unresolved. Who pays and who receives is a frequent point of litigation.[107]

3. Open source privileges

[104] P9_TA(2023)0236 Artificial Intelligence Act Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1.

[105] Quintais (2023) argues that the provision is impossible to comply with; Kretschmer et al. (2023) review such *ex ante* obligations as potentially problematic.

[106] Senftleben (2023). Council Directive 92/100/EEC of 19 November 1992 on rental right and lending right and on certain rights related to copyright in the field of intellectual, Official Journal of the European Communities 1992 L 346, 61.

[107] Kretschmer (2011); Peukert (2024).

Open source corpora and open source models have considerable advantages for the secure development and deployment of AI systems. Because of their transparency, open-source AI can potentially outperform closed AI systems, evidenced for example by the wide use of open source code in operating systems and security protocols. Models that disclose, even generally, their training sources show that repositories governed by open licences, such as Wikipedia or GitHub, are common sources of training data.[108]

The European Parliament's amendments to the proposed AI Act aim to provide extensive privileges to free and open-source AI components. Recital 12a states: "To foster the development and deployment of AI, especially by SMEs, start-ups, academic research but also by individuals, this Regulation should not apply to such free and open-source AI components except to the extent that they are placed on the market or put into service by a provider as part of a high-risk AI system or of an AI system that falls under Title II or IV of this Regulation." The wording is implemented under a new Art. 5(d).

As with Art. 28, the amendment may not survive the legislative process. However, exploring copyright liability privileges for the deployment of open source models remains an interesting avenue, for example by setting a time window for expedient correction.

For the established lifecycle of machine learning, we have shown that the mix of legal, technological and contractual opacity may lead to an undesirable allocation of licences and obligations. Training and deploying unlicensed models in the EU is currently risky, and will remain so for the foreseeable future. This makes it likely that practices in the EU will be moving towards a fully licensed AI copyright environment, regardless of the available exceptions. If model training needs to rely on permissions, the key question becomes where a suitable licence may be obtained and under what conditions. Market entry by European AI firms without the resources to access licensed corpora will become more difficult and costly.[109]

Is it for the public benefit to allow copyright works to be used, without permission, as training materials for machine learning? As a society, we don't know the answer yet, but the currently proposed copyright solutions may lead us into a fully licensed AI environment controlled by major rightholders and large AI firms. An alternative would be to take machine learning seriously as a general purpose technology.[110] Copyright law may not be able to solve the tensions between market entry, open source innovation and creator remuneration but it must try.

[108] Emanuilov and Margoni (2023).

[109] At the global level, the AI industry's emerging approach appears different for different media modes, i.e. for text, images, audio, video. It appears that infringement and potential disclosure issues will be more pertinent for music and visual content than words. For example, there are recent licensing deals reported between Google and Universal Music over recordings and "voices" (Financial Times 2023), or between Nvidia and Getty, Shutterstock and Adobe over images (Reuters 2023) while strong fair use claims under US law persist for literary productions.

[110] Kretschmer et al. (2023).

References

AirBnB (2023) Terms of service for European users. https://www.airbnb.co.uk/help/article/2908/terms-of-service#EU12. Accessed 30 Sept 2023

Arnold T, Tilton L, Berke A (2019) Visual style in two network era sitcoms. J Cult Anal. https://doi.org/10.22148/16.043

Birhane A, Kasirzadeh A, Leslie D et al (2023) Science in the age of large language models. Nat Rev Phys 5:277–280. https://doi.org/10.1038/s42254-023-00581-4

Boeing G, Waddell P (2017) New insights into rental housing markets across the United States: web scraping and analyzing Craigslist rental listings. J Plan Educ Res 37(1):457

Bommasani et al (2022) On the opportunities and risks of foundation models. Center for Research on Foundation Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI). arXiv:2108.07258

Brunstein D, Casamatta G, Giannoni S (2023) Using machine learning to estimate the heterogeneous impact of Airbnb on housing prices: evidence from Corsica (April 2, 2023). SSRN. https://doi.org/10.2139/ssrn.4407202

Buonocore T (2019) Man is to doctor as woman is to nurse: the gender bias of word embeddings. https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17. Published 8 March 2019

Burrow S (2021) The law of data scraping: a review of UK law on text and data mining. CREATe Working Paper 2021/2. Zenodo: https://doi.org/10.5281/zenodo.4635759

Campbell F (2019) Data scraping—what are the privacy implications? Privacy & Data Protection Journal 20(1) Oct/Nov 2019

Chalkidis I, Kampas D (2019) Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif Intell Law 27:171

Cambridge Consultants (2019) Use of AI in online content moderation. Ofcom Report, pp 51–52. https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation

CMA (2023) AI foundation models. Report by UK Competition and Markets Authority (CMA). https://www.gov.uk/government/publications/ai-foundation-models-initial-report. Published 18 Sept 2023

Cottman BH (2020) Converting PDF and Gutenberg document formats into text: natural language processing in production. https://towardsdatascience.com/natural-language-processing-in-production-converting-pdf-and-gutenberg-document-formats-into-text-9e7cd3046b33. Published 22 Aug 2020

Couldry N, Mejias UA (2019) Data colonialism: rethinking big data's relation to the contemporary subject. Telev New Media 20(4):336–349. https://doi.org/10.1177/1527476418796632

Craig CJ (2017) Globalizing user rights-talk: on copyright limits and rhetorical risks. Am Univ Int Law Rev 33:1

CSPLA, CNC and HADOPI (2020) Mission report: towards more effectiveness of copyright law on online content sharing platforms: overview of content recognition tools and possible ways forward (English version). Joint Report by CSPLA, CNC and HADOPI (January 2020)

Ducato R, Strowel A (2019) Limitations to text and data mining and consumer empowerment: making the case for a right to "machine legibility." IIC Int Rev Intellect Prop Competition Law 50:649. https://doi.org/10.1007/s40319-019-00833-w

Eben M, Erickson K, Kretschmer M et al (2023) Priorities for generative AI regulation in the UK: CREATe response to the Digital Regulation Cooperation Forum (DRCF). CREATe Working Paper 2023/8. Zenodo: https://doi.org/10.5281/zenodo.8319662

Eckart de Castilho et al (2018) A legal perspective on training models for natural language processing. LREC 2018

Elgammal A, Liu B, Elhoseiny M, Mazzone M (2017) CAN: creative adversarial networks generating "art" by learning about styles and deviating from style norms. extended version of a paper published on the eighth International Conference on Computational Creativity (ICCC), held in Atlanta, GA, 20–22 June 2017. arXiv:1706.07068v1

Elgendy M (2020) Deep learning for vision systems. Manning Publications, New York

Emanuilov I, Margoni T (2023) Forget me not: memorization in generative sequence models. Paper presented at 2023 EPIP conference

EUIPO (2020) Automated content recognition: discussion paper—phase 1 'existing technologies and their impact on IP'. https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory/documents/reports/2020_Automated_Content_Recognition/2020_Automated_Content_Recognition_Discussion_Paper_Full_EN.pdf.

Financial Times (2023) Google and Universal Music negotiate deal over AI 'deepfakes' (Anna Nicolaou and Madhumita Murgia). https://www.ft.com/content/6f022306-2f83-4da7-8066-51386e8fe63b. Published 8 Aug 2023

Flynn S et al (2020) Implementing user rights for research in the field of artificial intelligence: a call for international action. SSRN Electron J 42:393

Gatys LS, Ecker AS, Bethge M (2016) A neural algorithm of artistic style. J Vis 16:326. https://doi.org/10.1167/16.12.326

Geiger C, Frosio G, Bulayenko O (2018a) The exception for text and data mining (tdm) in the proposed Directive on Copyright in the Digital Single Market—legal aspects: in-depth analysis. (Policy Department for Citizens' Rights and Constitutional Affairs, Directorate General for Internal Policies of the Union) https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf

Geiger C, Frosio G, Bulayenko O (2018b) Text and data mining in the proposed Copyright Reform: making the EU ready for an age of big data? IIC Int Rev Intellect Prop Competition Law 49:814

Gold Z, Latonero M (2018) Robots welcome: ethical and legal considerations for web crawling and scraping. Wash J Law Tech Arts 13:275–281

Google (2019) Understanding searches better than ever before (blog by Pandu Nayak). https://blog.google/products/search/search-language-understanding-bert/. Published 25 Oct 2019

Guadamuz A, Cabell D (2014) Data mining in UK higher education institutions: law and policy. Queen Mary J Intellect Prop 4:3

Guérin J, Gibaru O, Thiery S, Nyiri E (2018) CNN features are also great at unsupervised classification. 8th international conference on computer science, engineering and application. Melbourne, Australia. https://doi.org/10.48550/arXiv.1707.01700

Handke C, Guibault L, Vallbé J-J (2021) Copyright's impact on data mining in academic research. Manag Decis Econ 42(8):1999–2016. https://doi.org/10.1002/mde.3354

Hillen J (2019) Web scraping for food price research. Br Food J 121(2):3350

Hirschey J (2014) Symbiotic relationships: pragmatic acceptance of data scraping. Berkeley Technology Law J 29:906

Ho HL (2015) The legal concept of evidence. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/evidence-legal/. First published 13 Nov 2015; substantive revision 8 Oct 2021

Hugenholtz PB (2016) Database directive. In: Dreier T, Hugenholtz PB (eds) Concise copyright law, 2nd edn. Kluwer, New York

Hughes A (2023) ChatGPT. BBC Science Focus. https://www.sciencefocus.com/future-technology/gpt-3. Published 25 Sept 2023

Jennings F, Yates J (2009) Scrapping over data: are the data scrapers' days numbered? JIPLP 4(2):120

Jurafsky D, Martin JH (2020) Speech and language processing. 3rd edn. https://web.stanford.edu/~jurafsky/slp3/

Kalehbasti R, Nikolenko L, Rezaei H (2021) Airbnb price prediction using machine learning and sentiment analysis. In: Holzinger A, Kieseberg P, Tjoa AM, Weippl E (eds) Machine learning and knowledge extraction. Lecture notes in computer science, vol 12844. Springer, Berlin. https://doi.org/10.1007/978-3-030-84060-0_11

Karpathy A et al (2014) Large-scale video classification with convolutional neural networks. IEEE conference on computer vision and pattern recognition. Columbus, OH, USA. https://doi.org/10.1109/CVPR.2014.223

Kavlakoglu E (2020) NLP vs. NLU vs. NLG: the differences between three natural language processing concepts. https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts. Published 12 Nov 2020

Keller P (2023) Generative AI and copyright: convergence of opt-outs? https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/. Published 23 Nov 2023

Kretschmer M (2011) Private copying and fair compensation: an empirical study of copyright levies in Europe. Intellectual Property Office Research Paper No. 2011/9. SSRN: https://doi.org/10.2139/ssrn.2710611

Kretschmer M, Kretschmer T, Peukert A, Peukert C (2023) The risks of risk-based AI regulation: taking liability seriously. CEPR Discussion Paper DP18517 (10 October 2023), also available via SSRN

Kumar S, Hosurmath M (2019) Multiclass image classification of yoga postures using Watson Studio and Deep Learning as a service. https://developer.ibm.com/technologies/artificial-intelligence/tutorials/image-preprocessing-for-computer-vision-usecases/

Levendowski A (2017) How copyright law can fix artificial intelligence's implicit bias problem. Wash Law Rev 93:579

Lunden I (2017) Airbnb eyes expansion with affiliate program for sites with 1M+ users, new API. https://techcrunch.com/2017/10/16/airbnb-eyes-expansion-with-affiliate-program-for-sites-with-1m-users-new-api/. Published 16 Oct 2017

Margoni T, Kretschmer M (2022) A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. GRUR Int 71(8):685–701. https://doi.org/10.1093/grurint/ikac054

Margoni T, Quintais JP, Schwemmer S (2022) Algorithmic propagation: do property rights in data increase bias in content moderation? (part I and II). Kluwer Copyright Blog. http://copyrightblog.kluweriplaw.com/2022/06/08/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-i/. Published 8/9 June 2022

Margoni T, Ducuing C, Schirru L (2023) Data property, data governance and common European data spaces. Computerrecht, Tijdschrift voor informatietechnolgie en recht, 3/2023/116, pp 202–211

Massimino B (2016) Accessing online data: web-crawling and information-scraping techniques to automate the assembly of research data. J Bus Logist 37(1):34

Miller AI (2019) The artist in the machine: the world of AI-powered creativity. MIT Press, Cambridge

Munzert S, Rubba C, Meißner P, Nyhuis D (2015) Automated data collection with R: a practical guide to web scraping and text mining. Wiley, New York

OpenAI (2018) Improving language understanding with unsupervised learning. https://openai.com/blog/language-unsupervised/. Published 11 June 2018

Otero BG (2021) Machine learning models under the copyright microscope: is EU copyright fit for purpose? GRUR Int 70:1043

Peukert C (2024) Copyright levies and cloud storage: ex-ante policy evaluation with a field experiment. Res Policy. https://doi.org/10.1016/j.respol.2023.104918

Przybyła P et al (2016) Text mining resources for the life sciences. Database 2016:baw45

Quintais JP (2023) Generative AI, copyright and the AI Act. Kluwer Copyright Blog. https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/. Published 9 May 2023

Reisner A (2023) Revealed: the authors whose pirated books are powering generative AI. The Atlantic (19 August 2023)

Reuters (2023) Adobe, Nvidia AI imagery systems aim to resolve copyright questions (Dawn Chmielewski and Stephen Nellis). https://www.reuters.com/technology/adobe-nvidia-ai-imagery-systems-aim-resolve-copyright-questions-2023-03-21/. Published 21 Mar 2023

Rosati E (2018) An EU text and data mining exception for the few: would it make sense? JIPLP 13:429

Sag M (2019) The new legal landscape for text mining and machine learning. J Copyr Soc USA 66:291

Sartor G, Loreggia A (2020) The impact of algorithms for online content filtering or moderation. European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs

Scassa T (2019) Ownership and control over publicly accessible platform data. Online Inf Rev 43(6):986

Schaul K, Chen S Y, Tiku N (2023) Inside the secret list of websites that make AI like ChatGPT sound smart. Washington Post, 19 April 2023

Schirru L, Margoni T (2023) Arts 3 and 4 of the CDSM Directive as regulatory interfaces: shaping contractual practices in the commercial scientific publishing and stock images sectors. Kluwer Copyright Blog (22 August 2023). https://copyrightblog.kluweriplaw.com/2023/08/22/arts-3-and-4-

of-the-cdsm-directive-as-regulatory-interfaces-shaping-contractual-practices-in-the-commercial-scientific-publishing-and-stock-images-sectors/
Seawright J, Gerring J (2008) Case selection techniques in case study research: a menu of qualitative and quantitative options. Political Res Q 61(2):94–308. https://doi.org/10.1177/1065912907313077
Seifert et al (2017) Visualizations of deep neural networks in computer vision: a survey. In: Cerquitelli T, Quercia D, Pasquale F (eds) Transparent data mining in big and small data. Springer, Berlin
Senftleben M (2023) Generative AI and author remuneration. Int Rev Intellect Prop Competition Law (IIC) 54:1535–1560. https://doi.org/10.1007/s40319-023-01399-4
Soper T (2020) 'OpenAI should be renamed ClosedAI': reaction to Microsoft's exclusive license of OpenAI's GPT-3. https://www.geekwire.com/2020/openai-renamed-closedai-reaction-microsofts-exclusive-license-openais-gpt-3/. Published 25 Sept 2020
Synodinou TE (2019) Lawfulness for users in European copyright law: acquis and perspectives. JIPITEC. 10:20
Tan T (2020) Evolution of language models: n-grams, word embeddings, attention & transformers. https://towardsdatascience.com/evolution-of-language-models-n-grams-word-embeddings-attention-transformers-a688151825d2. Published 19 March 2020
Ueno T (2021) The flexible copyright exception for 'non-enjoyment' purposes—recent amendment in Japan and its implication. GRUR Int 70:145
Vaswani T, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Advances in neural information processing systems. 30 (NIPS)
Zhang H, Nakamura T, Isohara T et al (2023) A review on machine unlearning. SN Comput Sci 4:337. https://doi.org/10.1007/s42979-023-01767-4