GRUR International, 73(2), 2024, 111–127

https://doi.org/10.1093/grurint/ikad140
Advance Access Publication Date: 13 January 2024
Article

ANDRES GUADAMUZ*

A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs

This article delves into the legal issues surrounding the use of copyrighted works in training artificial intelligence (AI). It examines two critical questions: firstly, whether accessing and analysing copyrighted works for AI training constitutes copyright infringement; and secondly, whether outputs generated by AI from these inputs infringe on copyright. The primary focus is on the United Kingdom's jurisdiction, with comparative analyses from EU law and a few US cases. The article aims to bring clarity to these multifaceted legal issues by thoroughly exploring the technicalities of machine learning and its implications for ongoing and future litigation. The hypothesis suggests that most inputs may fall under existing exceptions and limitations, but the outputs' legality may depend on individual case specifics. The conclusion reflects on the inevitable legal challenges that accompany technological advances, as happened in the past with file sharing and the early World Wide Web. The article highlights the current stage of AI technology and law, suggesting that legal precedents or legislative actions might eventually settle disputes. It also emphasizes the importance of finding technological solutions – like metadata standards or proactive initiatives – to balance copyright holders' rights with AI innovation. The article concludes by acknowledging the irreversible emergence of AI in our lives and the legal system's need to adapt, offering equitable solutions to copyright holders while fostering technological advancement.

I. Introduction

The process by which artificial intelligence (AI) 'learns' to do something, particularly to generate works that emulate human creativity, often relies on having access and analysing large numbers of those works, learning patterns to create its own versions. To do this, the computer program must have copies of works to analyse to produce new results.

So, two copyright questions arise from the above, one for the inputs, and one for the outputs. From the perspective of inputs, is the act of accessing, reading, preparing, analysing, and mining data an act of copyright infringement, and if so, are there any applicable defences? From the perspective of the outputs, could the copyright owner of one of the works used to teach the computer sue the maker for copyright infringement from the resulting works?

The answer to the input question could fall under the exclusive rights of the author, and as such one would need permission or a licence to perform some form of analysis on a body of copyright works. There is growing debate about whether this is full infringement, or whether there could be a limitation to copyright that would allow researchers to undertake some form of data mining to train algorithms. The answer to the output question is more complex, and it may be very specific to the particulars of each situation.

This article will explore both questions in detail, with a few caveats. While the main jurisdiction studied will be the United Kingdom, there will be some analysis of EU law, and a few US cases used for comparison. This subject is likely to attract litigation across different jurisdictions, so comparative analysis will be important. At the time of writing there are several ongoing copyright infringement lawsuits;[1] this paper will not deal with those at all – that will be a task for future research.

The objective of this article is to bring structure to the multifaceted legal issues and to offer clarity on both technical and legal aspects. It is probable that some ongoing litigation might hinge on the technicalities of machine learning. Therefore, before delving into the legal analysis, the article will aim to provide a comprehensive explanation of how machine learning operates, as this could potentially illuminate how future litigation might unfold. The working hypothesis of this study is that most inputs are likely to fall within the existing exceptions and

* Dr.; Reader in Intellectual Property Law, University of Sussex, United Kingdom. After posting a draft version, I received many comments and corrections. The ones reflected in the final draft come from Fabio Fumi, Nate Angell, Ryan Moulton, Cédric Manara, Alexandra Giannopoulou, and Ana Ramalho. I would also like to thank the anonymous peer-reviewers for their useful contributions, and editor Tian Lu for helping to put this version together. All remaining errors are exclusively my own. There is no conflict of interest in any of the technologies or companies mentioned in the article.

1 There is a growing list of cases, some of the most notable ones at the time of writing are: *Doe 1 et al v GitHub et al*, Case No 4:2022cv06823 (N.D. Cal.); *Andersen et al v Stability AI et al*, Case No 3:23-cv-00201 (N.D. Cal.); *Getty Images v Stability AI*, Case No 1:2023cv00135 (D. Del.); *Tremblay et al v OpenAI*, Case No 4:23-cv-03223 (N.D. Cal.); *Authors Guild v Open AI*, Case No 1:23-cv-829 (S.D.N.Y.); and *Getty Images v Sability AI* (England), Case IL-2023-000007.

limitations in the jurisdictions studied. However, there will inevitably be case law that probes the boundaries of such limitations. Regarding outputs, the resolution will probably be more dependent on individual cases, but some of these works might also fall under certain exceptions and limitations.

II. Inputs and outputs

1. Gathering inputs

The explosion in the sophistication of AI tools that we have experienced in recent years has come about because of two important developments, firstly the improvement and variety of machine-learning models, but most importantly the availability of large training datasets. While the definitions of artificial intelligence vary greatly,[2] the common element is the requirement for data as an input of whichever model is being used.

Some gathering can take place from legitimate sources, or some which may be paywalled. But the best source of data nowadays is the internet, particularly from publicly accessible websites and other online services. The legality of such scraping will be discussed later, but the web offers an easily accessible source of material for training in many types of work, including images,[3] music,[4] news,[5] text,[6] and social media posts,[7] just to name a few.

Take for example a large language model such as Generative Pre-trained Transformer (GPT). Version 3 of this model was trained with 499 billion tokens[8] of data. A token is a sequence of characters that are often found together in the corpus of training text, usually made up of four letters or numbers. For example, the digits 1234567890 consist of four tokens 123, 45, 678, 90.[9] Some words map to one token, while some do not.[10] OpenAI has not released a lot of information on GPT-4, but GPT-3 was trained using mostly online sources, the vast majority from web crawl searches, some coming from books, but some of coming from other curated sources such as Wikipedia.[11] While OpenAI does not specify it, some of these sources were collected by OpenAI itself, but in other instances they used datasets created by others, such as the case of 16% of the training data coming from independent sources, particularly the book data, named Books1 and Books2.[12] There has been some online speculation[13] as to which are the sources for these two datasets, one comprising 12 billion tokens and the other 55 billion tokens respectively. It is almost certain that Books1 is made up of public domain books made available through Project Guttenberg,[14] while the content of Books2 remains a mystery.[15]

Another popular dataset used in training language models is The Pile,[16] which contains 22 other smaller datasets with content such as web crawls (Common Crawl and OpenWebText), PubMed, ArXiv articles, Wikipedia, the USPTO, Project Gutenberg, and Books3[17] (which is a dataset that contains fiction and non-fiction books collected in a torrent tracker called Bibliotik). Books3 appears to consist mostly of pirated copies of books.

Other datasets are built using content from data gathered from specific online sources. Take for example the WebVid dataset, which consists of 10 million video clips and accompanying description, taken from stock image websites.[18] This dataset is remarkable because, while it is created by researchers at Oxford University, it is used by Meta in the training of their own video generation model.[19]

Another dataset that has been subject to a great amount of scrutiny is the Large-scale Artificial Intelligence Open Network (LAION).[20] This is the largest image-text pair dataset at the time of writing, comprising 5.8 billion single entries. LAION was compiled using a web crawler that searched the public internet for HTML code containing both an image link element and an alt-text description of the image. These two elements make up the main part of the dataset, with other elements such as width, height, licence, and the

2 For some definitions, see: Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (CRC Press 2004) 423; Douglas R Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid* (20th anniversary edn, Penguin 2000) 601; and Stuart J Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (4th edn, Pearson 2020) 45; and art 3(1) of the 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, COM/2021/206 Final.

3 An early example used by many researchers is ImageNet, see: <https://www.image-net.org/> accessed 25 November 2023.

4 <https://freemusicarchive.org/> accessed 25 November 2023.

5 <https://www.kaggle.com/datasets/therohk/global-news-week> accessed 25 November 2023.

6 Such as a dataset of Amazon reviews, see: <https://www.cs.jhu.edu/~mdredze/datasets/sentiment/> accessed 25 November 2023.

7 A collection of different social media sources: <http://datasets.syr.edu/> accessed 25 November 2023.

8 Chuan Li, 'OpenAI's GPT-3 Language Model: A Technical Overview' (*Lambda Blog*, 3 June 2020) <https://lambdalabs.com/blog/demystifying-gpt-3/> accessed 25 November 2023.

9 Raf Khan, 'What are tokens and how to count them?' (*OpenAI blog*, 2023) <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them> accessed 25 November 2023.

10 You can check how many tokens can be found in any given text, see: <https://beta.openai.com/tokenizer> accessed 25 November 2023.

11 Tom B Brown and others, 'Language Models Are Few-Shot Learners' (*arXiv*, 22 July 2020) <http://arxiv.org/abs/2005.14165> accessed 25 November 2023.

12 ibid 9.

13 See for example this 2021 thread in Hacker News: <https://news.ycombinator.com/item?id=26308339> accessed 25 November 2023.

14 Martin Gerlach and Francesc Font-Clos, 'A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics' (*arXiv*, 19 December 2018) <http://arxiv.org/abs/1812.08092> accessed 25 November 2023.

15 See: Gregory Roberts, 'AI Training Datasets: the Books1+Books2 that Big AI eats for breakfast' (*Musings on Freedom Blog*, 14 December 2022) <https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/#Books1_Books2_15> accessed 25 November 2023.

16 <https://pile.eleuther.ai/> accessed 25 November 2023.

17 <https://huggingface.co/datasets/the_pile_books3> accessed 25 November 2023.

18 <https://m-bain.github.io/webvid-dataset/> accessed 25 November 2023.

19 Uriel Singer and others, 'Make-A-Video: Text-to-Video Generation without Text-Video Data' (*arXiv*, 29 September 2022) <http://arxiv.org/abs/2209.14792>. accessed 25 November 2023

20 Romain Beaumont, 'LAION-5b: A new era of open large-scale multi-modal datasets' (*LAION Blog*, 31 March 2022) <https://laion.ai/blog/laion-5b/> accessed 25 November 2023.

likelihood that the image may be 'unsafe', as well as providing a similarity score.[21]

However, some companies working on AI are not as transparent with their datasets. Midjourney, one of the early successful companies in the space of image generation, has not disclosed which datasets it uses.[22] Similarly, OpenAI does not always disclose the datasets used in its products.

One thing that is evident from the above is that the nature of collected data and the method of collection will be as varied as the number of datasets. Some datasets contain full works, such as images, video, text, and music. Some contain metadata, while some contain links to data. That means the nature of collection will be varied. Some will require a copy of a work to be made and kept in the dataset, while others will require just a temporary copy for the purpose of extracting data. A good example of this is the aforementioned LAION 5b dataset; while the final corpus consists of links, the images are processed for the extraction of safety and similarity data. LAION describes the process as downloading images for analysing in batches, extracting the required information, and then discarding the data.[23]

There is a final element to discuss in the input phase, and that is a process that precedes the training of a model. This is known as data preparation,[24] which can be described as the process of transforming raw data into formats that can be analysed using machine learning algorithms. The actual preparation will of course depend on the type of training, as well as the data itself, but it usually requires arranging data into a format that can be analysed and used by the process of machine learning, and can refer to the addition, deletion, or transformation of training set data.[25] While this could be performed manually in a small dataset, larger ones require automated tools for preparation. This could also require the copying and storage of data while it is being prepared.

2. Generating outputs: machine learning models

The generation of inputs in the shape of datasets as described above has several objectives, including data analysis, data storage, data knowledge extraction, and other similar functions.[26] However, when it comes to AI, data is used specifically to train models that will then be used in one of its functions. This is normally referred to as machine learning, where algorithms enable computers to learn from data and even improve themselves without being explicitly programmed.[27] The system improves as it is presented with more data.

The biggest conflict at present is in what is referred to as *generative AI*.[28] While the concept has become a bit of a buzzword, generative AI is generally understood as content that is generated from scratch using some form of machine learning algorithm.[29] A generative AI therefore can take a prompt and produce an entirely new work.[30] How is this possible? There is a common misconception that a generative operation is akin to putting together a collage of pre-existing images,[31] but this couldn't be further from the truth. Generative algorithms are a class of algorithms used in machine learning that can generate new outputs that are like their training dataset.[32] These algorithms typically work by building a model of the training data. In this context, a model is defined as a mathematical representation of a real-world process that is trained using a dataset. This can be used to make predictions or decisions without being explicitly programmed to perform the task,[33] and then using that model to generate new data samples that resemble the ones in the training dataset. The 'resemble' part of the explanation is important for the purposes of this work because the outputs are similar, not the same. The main idea behind creative AI is to train a system in a way that it can generate outputs that statistically resemble their training data. In other words, to generate poetry, you train the AI with poetry; if you want it to generate paintings, you train it with paintings.

There are many different types of generative algorithms, the most common include variational autoencoders (VAEs),[34] autoregressive models,[35] generative adversarial networks (GANs),[36] and diffusion

21 For example, entry ID 1,996,868,005,430 contains a link image element 'https://images-na.ssl-images-amazon.com/images/I/5113P2xd-MGL.jpg', as well as the description 'Logans Run Import Anglais'.

22 Janus Rose, 'Inside Midjourney, The Generative Art AI That Rivals DALL-E' (*Vice*, 18 July 2022) <https://www.vice.com/en/article/wxn5wn/inside-midjourney-the-generative-art-ai-that-rivals-dall-e> accessed 25 November 2023.

23 Beaumont (n 20).

24 Jason Brownlee, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python* (Machine Learning Mastery 2020).

25 Max Kuhn and Kjell Johnson, *Applied Predictive Modeling* (Springer 2018) 27.

26 See for example: Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, 'From Data Mining to Knowledge Discovery in Databases' (1996) 17 AI Magazine 37.

27 Michael Jordan and TM Mitchell, 'Machine Learning: Trends, Perspectives, and Prospects' (2015) 349 Science 255.

28 Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan, 'ChatGPT Is Not All You Need. A State of the Art Review of Large Generative AI Models' (*arXiv*, 11 January 2023) <http://arxiv.org/abs/2301.04655> accessed 25 November 2023.

29 Chiara Longoni and others, 'News from Generative Artificial Intelligence Is Believed Less' 2022 ACM Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery 2022) <https://doi.org/10.1145/3531146.3533077> accessed 25 November 2023. See also Tijn van der Zant, Matthijs Kouw and Lambert Schomaker, 'Generative Artificial Intelligence' in Vincent C Müller (ed), *Philosophy and Theory of Artificial Intelligence* (Springer 2013).

30 For example, 'a mice and a squirrel on their mobile phones': <https://s.mj.run/qE_U8-QFmpU> accessed 25 November 2023.

31 See, for example, this incorrect explanation: <https://twitter.com/ZedEdge/status/1594133323710070785> accessed 25 November 2023.

32 Andrey Karpathy and others, 'Generative Models' (*OpenAI Blog*, 16 June 2016) <https://openai.com/blog/generative-models> accessed 25 November 2023.

33 There are many types of machine learning models, such as decision trees, random forests, neural networks, and support vector machines. See: Yang Lu, 'Artificial Intelligence: A Survey on Evolution, Models, Applications and Future Trends' (2019) 6 Journal of Management Analytics 1.

34 Lucas Pinheiro Cinelli and others, *Variational Methods for Machine Learning with Applications to Deep Networks* (Springer Nature 2021) 35.

35 McClain Thiel, 'An Intuitive Introduction to Deep Autoregressive Networks' (*Machine Learning Berkley Blog*,16 August 2020) <https://ml.berkeley.edu/blog/posts/AR_intro/> accessed 25 November 2023.

36 Ian J Goodfellow and others, 'Generative Adversarial Networks' (*arXiv*, 10 June 2014) <http://arxiv.org/abs/1406.2661> accessed 25 November 2023.

models.[37] These algorithms can be used for a wide range of tasks, such as image generation, data augmentation, and anomaly detection.[38]

The most successful recent examples of generative AI in images, such as Imagen, DALL·E 2, Stable Diffusion, and Midjourney, use the diffusion model, which reportedly produces superior results.[39] Diffusion works by taking an input, for example a photograph, and then corrupting it by adding noise to it; the training takes place by teaching a neural network to put it back together by reversing the corruption process.[40]

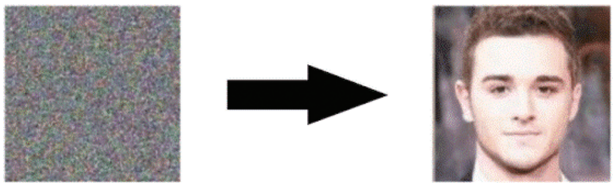

Image 1: *Diffusion model reconstructing an image*[41]

This is useful in machine learning because the algorithm is learning to put an image back together, which means that it can better 'remember' what something looks like by turning it into noise and then reversing the process (Image 1). Just as with other generative models, the output is not an exact replica of the training data, it is a statistical approximation of it.

But how does a diffusion model know what an image looks like when prompted? One of the most impressive aspects of the AI tools that have become popular in the last year is precisely that the machine can understand what one means when typing 'a cat in the style of van Gogh'. The AI must be able to understand what a cat looks like, as well as know how to apply the unique van Gogh style to it to generate a new image. So, it needs not only to have been trained on cats and van Gogh, but it must also understand what the words mean and match them with the images in the dataset.

This is where another machine learning model comes in. Contrastive Language-Image Pre-training (CLIP) combines the strengths of both convolutional neural networks (CNNs)[42] and large language models (LLMs), and is designed to improve the performance of AI models on a wide range of tasks involving both language and images. The model is trained using a large dataset of images and their corresponding text descriptions, and it learns to understand the relationship between language and images.[43]

So, AI tools use a combination of a diffusion model trained on reconstructing images and natural language models that understand words used to describe an image.[44] If I type the word 'cat' into Stable Diffusion, it understands because it was trained on thousands of reference images of cats and their accompanying descriptions, and it can build a cat from scratch as it 'remembers' what a cat looks like from having diffused it. This combination of language and image models allow me to type 'a llama by Gustav Klimt' with DALL-E, and it will be able to comply and produce an output.[45]

There is another important element involved in generative models, and this is called latent space.[46] In order to train a model with millions – and sometimes billions – of single data points, it would be inefficient to treat every data point in the same way as there could be clusters of similar images that do not need to be memorised. If we are thinking about images, we may not have to look at every single cat picture; it may suffice to cluster data that are similar. In many ways our brain works like this – you don't need to know every single shape or size of a dog to recognise one on the street: at some point you learn that anything from a Chihuahua to a Labrador are dogs.[47] So a latent space is a lower-dimensional space where each point represents a different 'embedding' or 'code' for the original data; the idea is to find a more compact representation of the data. The wording here is precise for a reason: notice that latent space does not contain copies of works, but clustered and statistical representations of works.

So, imagine data as a room: you would put the cat representations in the same space, the dog representations in another space; and under those divisions, angora cats, Siamese cats, etc. Latent space allows the classification and clustering of similar data and word representations. It is used in generative models where the goal is to learn a representation of the data that can help generate new samples that are like the ones in the training set (Image 2). This is valuable because it helps to understand why the inputs are needed: they're not copied, the accumulated representation of items is extracted.[48] This is vital in understanding the issue of generating outputs. Training models do not hold every piece of data in their training, they hold data representations clustered into similar works.

This is perhaps one of the most important parts of machine learning models for copyright purposes, and one that is often misrepresented. There is a common

37 Jonathan Ho and Chitwan Saharia, 'High Fidelity Image Generation Using Diffusion Models' (*Google Research Blog*, 16 July 2021) <https://ai.googleblog.com/2021/07/high-fidelity-image-generation-using.html> accessed 25 November 2023.

38 Karpathy (n 32).

39 Parfulla Dhariwal and Alexander Nichol, 'Diffusion Models Beat GANs on Image Synthesis', Part of (2021) 'Advances in Neural Information Processing Systems' 34 (*NeurIPS*, 2021) <https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html> accessed 25 November 2023.

40 Ryan O'Connor, 'Introduction to Diffusion Models for Machine Learning' (*AssemblyAI Blog*, 22 May 2022) <https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/> accessed 25 November 2023.

41 ibid.

42 Jiuxiang Gu and others, 'Recent Advances in Convolutional Neural Networks (2018) 77 Pattern Recognition 354.

43 ibid.

44 Ian Stenbit, 'A walk through latent space with Stable Diffusion' (28 September 2022) <https://keras.io/examples/generative/random_walks_with_stable_diffusion/> accessed 25 November 2023.

45 See one here: <https://labs.openai.com/s/Wh3iFd72wvg2ENMYoiMBLDhm> accessed 25 November 2023.

46 Ekin Tiu, 'Understanding Latent Space in Machine Learning' (*Towards Data Science*, 4 February 2020) <https://towardsdatascience.com/understanding-latent-space-in-machine-learning-de5a7c687d8d> accessed 25 November 2023.

47 Panagiotis Antoniades, 'Latent Space in Deep Learning' (*Baeldung Blog*, 5 November 2022) <https://www.baeldung.com/cs/dl-latent-space> accessed 25 November 2023.

48 For a comparison of latent space in different models, see: Andrea Asperti and Valerio Tonelli, 'Comparing the Latent Space of Generative Models' (2023) 35 Neural Computing and Applications 3155.

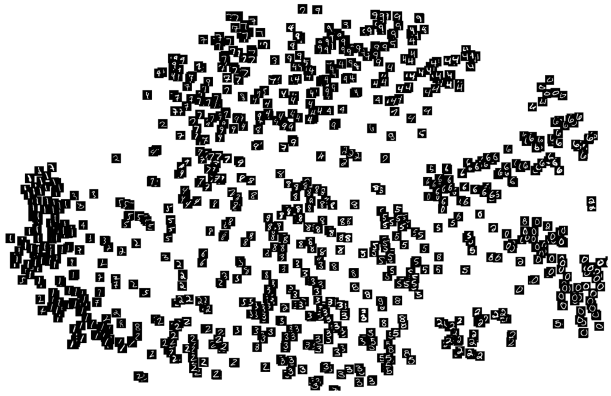

Image 2: *A latent space of hand-drawn numbers from 0-9, they're clustered by similarity*[49]

misconception that a model is just a dataset that contains a copy of every single entry in the dataset. This is where the distinction between the data in the inputs and the models themselves becomes relevant. We can have a dataset such as the Guttenberg corpus of public domain books, and we could train a large language model with only that dataset. The trained model does not contain a copy of the books; at the risk of over-simplifying things, such a model contains statistics, the likelihood of a token following another.[50] Obviously this is more complicated than just mere statistics. There is a complex interaction of neural networks, and we could go into more detail regarding concepts such as transformers,[51] which we do not need to cover now. But the takeaway is that models do not contain copies of works. For legal purposes, they are not even derivatives of one specific work in the dataset; the idea of large datasets is precisely that a model is not based on one individual work. This is evidenced by the size of trained models. While some are relatively large,[52] they run in gigabytes, not tera or petabytes, and cannot possibly contain all the works in the training data – they are mostly statistical data.

This brings us to a final characteristic of models, their persistence. Once trained, the model does not need the data it was trained on (which is another common misconception). A trained model does not need to be updated, it can be run independently, and in theory it can subsist forever if it is hosted or stored. A trained model then consists mostly of 'weights'.[53] These are parameters in a neural network that are the result of translating the inputs present in the training data into data points understandable by the specific model. So, when asking an AI tool to generate a work, it does not go back to the training data, it is working from a leaner and much more concise set of abstract concepts. But the result is the same – this is not a copy of a work, and it is not a derivative of a specific work. This could prove important for copyright liability issues, as trained models are in theory capable of being in existence forever once trained, even if one cuts access to all its training data. In some ways, we could look at models in the same way that we view some peer-to-peer networks, i.e. almost impossible to bring down.

III. Copyright and inputs

The models described in the previous section require large amount of training data. The more data available the better the models, which translate into more accurate weights, which can then be used to produce better outputs.[54]

The question arises as to whether the use of data in the training of such models is infringing copyright. As described above, data collection almost certainly will require making a copy of the data, be it in the shape of text, images, music, paintings, portraits, etc. From a technical perspective, whichever method one is using to train and teach the AI to do something, this will require accessing and reading the data. This will be stored and then analysed, often repeatedly, to extract information, produce statistical analysis, and produce outputs, all depending on the model. For example, to train the language model in GPT-3 (and its precursors), researchers scraped text from websites with a web crawler which extracted the HTML code from the target web pages, and then the text was extracted from that code to produce a full dataset.[55] This means an actual copy of the text from the webpage was made at some point. Some models could use a temporary copy.[56]

The question of whether the copying and use of this data could be infringing copyright will depend on a variety of factors. We will concentrate on two main steps: the possible copyright infringement in the mining of data and training of models; and, if there is infringement, the existence of possible exceptions and limitations.

1. Is there infringement?

The analysis of the existence of copyright in the training data is important because not all data has copyright.

49 Julien Despois, 'Latent space visualization' (*Hackernoon*, 10 July 2017) <https://hackernoon.com/latent-space-visualization-deep-learning-bits-2-bd09a46920df> accessed 25 November 2023.

50 For a comprehensive yet understandable explanation of how a language model like GPT works, see: Stephen Wolfram, 'What Is ChatGPT Doing … and Why Does It Work? (*Stephen Wolfram Writings*, 14 February 2023) <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/> accessed 25 November 2023.

51 A transformer is a type of model that processes information differently than other models. Instead of looking at data in a specific order (like reading words in a sentence one by one), it looks at all the data at once and figures out the relationships between different parts; see: Thomas Wolf and others, 'Transformers: State-of-the-Art Natural Language Processing' (2020) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (*ACL Anthology*, October 2020) <https://aclanthology.org/2020.emnlp-demos.6> accessed 25 November 2023.

52 Take for example Stable Diffusion 2.1, which initially runs at around 5.2GB, while the language model LLAMA 65B runs at 31.2GB.

53 Kurava S Ganesh, 'What's The Role Of Weights And Bias In a Neural Network?' (*Towards Data Science*, 24 July 2020) <https://towardsdatascience.com/whats-the-role-of-weights-and-bias-in-a-neural-network-4cf7e9888a0f> accessed 25 November 2023.

54 See: Christopher Bowles and others, 'GAN Augmentation: Augmenting Training Data Using Generative Adversarial Networks' (*arXiv*, 25 October 2018) <http://arxiv.org/abs/1810.10863> accessed 25 November 2023.

55 Alec Radford and others, 'Language Models are Unsupervised Multitask Learners' (2018) OpenAI Research Paper <https://bit.ly/3m-fceXg> accessed 25 November 2023.

56 Some GANs can use temporary data from various data sources, see: Hui Qu and others, 'Learn Distributed GAN with Temporary Discriminators' (*arXiv*, 17 July 2020) <http://arxiv.org/abs/2007.09221> accessed 25 November 2023.

In the broadest sense, data are comprised of individual facts, numbers, statistics, or items of information that can be collected.[57] Some of these items can be works that are capable of copyright protection, such as text, photographs, sound recordings, music, artworks, etc. But facts, numbers, and information may not always be protected.[58]

For copyright to subsist, the work must be protected subject matter and needs to be original, that is, the author's own intellectual creation.[59] It is evident that this covers individual works such as books, music and art that could be included in a database, but this may not be the case with just raw information. Large data gathering cannot be said to meet this requirement as it is often done automatically. This means some data will be in the public domain due to lack of protection, such as raw data, numbers, bits of information and any type of similar facts that are collected without any human intervention and so may not be protected.[60] Similarly, the gathering of works that are already in the public domain for inclusion on a database would not be copyright infringement.

Given the possible difficulty in trying to discern ahead of time if a dataset is subject to copyright protection, data mining operations can bypass most legal issues by using data from legitimate sources, either using public domain works, or data gathered with the authorisation of the owner, pre-empting any possible infringement questions. Other legitimate sources could be open access datasets or works released under permissible some-rights-reserved licences such as Creative Commons,[61] or open-source software licences. So, developers wanting to train their models will have a varied number of legitimate and non-infringing data sources at their disposal.

However, these may not be enough. If one were to rely only on public domain works not only would a lot of valuable data be excluded, but it would also be heavily biased towards a few cultures that have had their cultural works archived.[62] This is where copyright comes into play. Assuming a developer wanted to gather the text of every available website on the internet, or wanted to mine as many images as possible from the public web, would such collection be infringing?

In principle, any collection of protected works without permission will infringe copyright if it involves one of the exclusive rights of the author, such as reproduction, distribution, adaptation, communication to the public, etc.[63] So, copying a video without authorisation, such as WebVid, would be an infringement. If this is the case, developers and researchers will have to rely on an exception to copyright, something we will deal with later.

But it must be stressed here that the collection must be performing one of the exclusive rights of the author. The copying is an easy one, as it infringes the right of reproduction. But what if the work is not being copied? This is what happens with many of the datasets described in previous sections, particularly metadata ones where a work is not actually copied but rather some information about a work is extracted – like, for example, the musical attributes of songs by Billie Eilish,[64] or the song data from every Taylor Swift album.[65]

As described earlier, LAION is an interesting example of a dataset that does not contain copies of works, but rather extracts image links and text descriptions. While some temporary copying is undertaken during the collection of links to extract a similarity score, the dataset does not keep those temporary reproductions. One could argue that placing the links on a database could be a communication to the public,[66] and while many images may have been shared by their owners, some others may not. It is unlikely that these links could be a communication to the public because they are not being made public in the strictest sense. This is a dataset with over 5 billion other links, so entries in the database are not easy to access – it takes a lot of effort to be able to download the entire dataset, find a specific link, and access it. There are some browsers capable of sifting and searching through the links[67] and collecting the images, but one could argue that this is the same as browsing the web, which stretches the definition of making something accessible to the public. Moreover, there is a considerable corpus of case law that explains what communication to the public is.[68] It is not the remit of this paper to explain this just for one dataset, but an analysis of the existing case law leads one to believe that links to a potentially infringing copy of a work need not necessarily be infringing, and most importantly, more often than not the link will lead to a copy that has been made public by the owner.[69]

If the dataset is not infringing, could the actual training and creation of a model be infringing copyright?

57 The Wikipedia definition is as good as any: <https://en.wikipedia.org/wiki/Data> accessed 25 November 2023.

58 For example, in the US facts and data are not protected as per *Feist Publications Inc v. Rural Telephone Service Company Inc* 499 U.S. 340; see also Pamela Samuelson, 'Copyright Law and Electronic Compilations of Data' (1992) 35 Communications of the ACM 27.

59 Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* ECLI:EU:C:2009:465, para 37.

60 As of writing, the intellectual creation originality standards still also apply in the UK since *SAS Institute Inc v World Programming Ltd* [2013] EWHC 69; it is unlikely to change in the near future as per *Tunein Inc v Warner Music UK Ltd & Anor* [2021] EWCA Civ 441.

61 See for example the Multimedia Commons, a collection of audio and visual works licensed with Creative Commons <https://registry.opendata.aws/multimedia-commons/> accessed 25 November 2023.

62 For more on bias in training data, see: Sunyam Bagga and Andrew Piper, 'Measuring the Effects of Bias in Training Data for Literary Classification' (2002) Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature 74-84.

63 s 19 CDPA.

64 <https://data.world/priyankad0993/billie-elish-music> accessed 25 November 2023.

65 <https://data.world/promptcloud/taylor-swift-song-data-from-all-the-albums> accessed 25 November 2023.

66 See Justin Koo, *The Right of Communication to the Public in EU Copyright Law* (Bloomsbury Publishing 2019).

67 One such browser can be found here: <https://haveibeentrained.com/> accessed 25 November 2023.

68 Amongst others, Case C-466/12 *Nils Svensson and Others v Retriever Sverige AB* ECLI:EU:C:2014:76; Case C-160/15 *GS Media BV v Sanoma Media Netherlands BV and Others* ECLI:EU:C:2016:644; Case C-610/15 *Stichting Brein v Ziggo BV* ECLI:EU:C:2017:456; and Case C-527/15 *Stichting Brein v Filmspeler* ECLI:EU:C:2017:300.

69 For an excellent discussion of the recent cases, see: Jane C Ginsburg, 'The Court of Justice of the European Union Creates an EU law of liability for facilitation of copyright infringement: Observations on Brein v. Filmspeler [C-527/15] (2017) and Brein v. Ziggo [C-610/15]' (2017) Columbia Law and Economics Working Paper No 572 <https://ssrn.com/abstract=3024302> accessed 25 November 2023.

As explained in a previous section, a copy of the data must be available in some form for preparation and data extraction. So, there could be infringement if this reproduction is unauthorised.

After the model has been trained, the copied data are not needed anymore, unless one is training other models on the same data. A trained machine learning model contains the set of parameters that have been learned from training data. These parameters are used to make predictions on new data, and the quality of the predictions is determined by how well the model has learned the relationship between the input data and the corresponding output labels.[70] In other words, the resulting trained model does not contain copies of the dataset, but highly compressed information in a latent space.

To better illustrate this, we can follow how a model such as Stable Diffusion is trained.[71] Version 2.0 is trained on images that have been gathered using the LAION-5b dataset. You might recall that the training takes images and encodes them into latent space representations, a process that clusters images and words. These representations are processed with noise, and then de-noised, because this is a diffusion model. The result is a model that contains data points of things present in the dataset, which contains representations of cats, dogs, buildings, monuments, and people. However these are not copies in the legal sense; this is lossy data.

Are these representations potentially adaptations of the images? It is difficult to see how they could be adaptations of a single work as the very process of training is an exercise into breaking things apart, clustering, putting things that are similar together, and then passing them through a noise filter. This is not a translation, there is practically nothing of the original left in the model (with some exceptions that we will cover later). Moreover, the actual content of the trained model is not browsable in the same way one can browse the contents of a dataset, and data scientists tend to evaluate the results of the training using statistical tools.[72]

2. Exceptions and limitations

Not all datasets infringe copyright. However, there may be some cases where either the collection or the model training may infringe an exclusive right of the author. In those cases, the next question is whether there may be an exception to copyright that applies.

There is an exception that could apply depending on the technical nature of the training. As has been described above, it is not necessary for a trained model to contain copies of works in its final form, although there could be some copies made in a transitory manner during the training process. So many models are trained with the making of a temporary copy, and this is turned into an abstract version of the work in latent space, which is then used to produce outputs. In the UK, Sec. 28A of the CDPA[73] states that copyright is not infringed in the making of a temporary copy which is transient or incidental, if this copy is an integral part of a technological process, and the sole purpose is to either enable a transmission of the work, or for a lawful use, and the temporary copy does not have independent economic significance.

It could be argued that the training of an AI model does not fulfil all these requirements, but it is likely to be something that will be argued in court by future defendants in copyright litigation. Making a copy only for training is transient, but it could be said that it is not incidental, it is required to extract information, and without these copies there is no trained model. The copy is also part of a technological process, but it may not be considered a lawful use. Similarly, one could argue that the resulting model does have economic significance, but specific copies do not; a model such as Stable Diffusion can be trained with billions of images, each individual copy used in training may not count as having 'independent economic significance'. Take a language model that has been trained with this article after it has been published and made available online. If you remove the article from the training data, the model works just as well. This is because what matters for a model is not any individual work, but the extraction of accumulated information. In the case of text, the value is the analysis of billions of tokens and what matters is not which specific work is present, but that the number of works is large and varied. This highlights a fundamental challenge in copyright law concerning AI: discerning the individual value of works in a vast dataset versus the collective value extracted from the aggregation of these works.

The case law dealing with temporary copies may help to elucidate this point. In *Infopaq II*,[74] the Court of Justice of the European Union (CJEU) found that an electronic cutting service had made temporary copies of protected works carried out during a data capture process, and that this met the requirements of a temporary copy set out in Art. 5(1) of the InfoSoc Directive.[75] In the long-running case of *Meltwater*,[76] the question of whether making transient copies while viewing the internet was under discussion, the CJEU also found that copies made by a user while browsing the web and in the course of viewing a website also satisfied the requirements of Art. 5(1). While none of these cases are directly applicable to the training of a model, they give an indication that courts will be willing to give a broad interpretation to the existing requirement of what is a temporary copy.[77] When faced

70 Examples of how trained models work can be found here: Eli Stevens, Luca Antiga and Thomas Viehmann, *Deep Learning with PyTorch* (Simon and Schuster 2020) 83-97.

71 See here: <https://huggingface.co/stabilityai/stable-diffusion-2-base> accessed 25 November 2023.

72 See one in action here: <https://github.com/CompVis/stable-diffusion> accessed 25 November 2023.

73 This is derived from art 5(1) of the 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 [2001] OJ L167/10-19 (InfoSoc Directive).

74 Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening (Infopaq II)* ECLI:EU:C:2012:16.

75 Above n 73.

76 *Newspaper Licensing Agency Ltd v Meltwater Holding BV* C-360/13 [2014] AC 1438 (CJ).

77 For some scenarios that could meet the transient copy exception, see: Enrico Bonadio, Luke McDonagh and Plamen Dinev, 'Can Artificial Intelligence Infringe Copyright? Some Reflections' in Ryan Abbot, *Research Handbook on Intellectual Property and Artificial Intelligence* (Edward Elgar 2022).

with a claim of copies being transient, it is possible that the courts will have to look in detail at the technical features of training a model, and it will be interesting to see if future litigation argues this point.

The other main exception applicable to the training of a machine learning model is the existence of a limitation that deals specifically with the collection of the data itself. The earliest source of an exception for training an AI can be found in the United States in the shape of *Google Books*.[78] The case originated in 2004, when Google announced a new service called Google Print (later renamed Google Book Search). Google entered into an agreement with several libraries in the US and the UK which allowed them to digitise out-of-print books and make them available to anyone searching for that title. The books were to be offered either in their entirety, or in preview mode, meaning that only some pages would be accessible. Google's goal was to digitize 15 million books within a decade.[79]

Some authors responded negatively to Google's plans and, in 2005, two separate lawsuits were brought against the search engine giant by the Association of American Publishers and the Authors Guild. Both complaints were similar in scope and alleged that Google was engaging in copyright infringement by digitally reproducing the plaintiff's works for commercial gain, and then publicly distributing and displaying copies of those works. Google argued their actions fell under the *fair use* doctrine. The cases were initially settled out of court in 2008,[80] and this agreement contained several concessions from both parties that would allow Google Books to go ahead but offering only a few snippets. Authors could also monetise their books through Google.[81] Given growing criticism about the settlement from academics and authors,[82] the settlement was abandoned, and the trial continued. At the District Court level[83] the judge summarily dismissed the case, ruling that Google's actions amounted to fair use. The decision was appealed by the Authors Guild, and the Second Circuit affirmed the District Court's decision agreeing that Google's scanning was fair use. The transformative nature of the scanning played a big part in the decision,[84] as did the fact that the copying would not affect the market for book sales online because the purpose of the Google database was to make the works available to libraries, as well as to provide snippets in search results.[85]

While Google Books does not deal specifically with AI, it is similar in many ways to what happens in most machine learning training, i.e. there is copying of large amounts of works to produce something different. Commentators are divided on whether data mining falls under fair use in the US, and while some argue that it is fair use unequivocally,[86] others are less sure,[87] and see both scenarios as problematic.[88] We may have to wait for future litigation regarding machine learning specifically to get a final answer on this. At the time of writing, there are several ongoing cases that could eventually lead to a more complete answer.[89]

In other jurisdictions there are applicable exceptions for text and data mining (TDM), and a survey published in the journal Science lists several countries that contain an exception of some form.[90] One of the first countries to include a TDM exception in its law was the UK, as a result of a specific recommendation made by the Hargreaves Review of Intellectual Property,[91] which prompted the government to adopt a fair dealing exception to copyright for text and data mining for non-commercial purposes. It is important to point out here that the initial purpose of this exception (and something that is common in other similar fair dealing provisions) was that TDM was seen as something that would benefit scientific research in general. The review uses malaria research as an example where copyright law stifles innovation by making it difficult for scientists to analyse a large amount of data that could produce life-saving medicines.[92]

The resulting reform[93] allows for the creation of copies for the purpose of text and data analysis, if it is done 'for the sole purpose of research for a non-commercial purpose'.[94] It must also provide sufficient acknowledgement to the copied work, unless this is not possible for practical reasons.[95] This exception could also be seen as complimentary to other existing fair dealing provisions for scientific research pre-dating the 2014 reform.[96]

This exception would cover several uses by researchers, a good example is WebVid, where full copies of video clips have been reproduced and used in the

78 *Authors Guild v Google, Inc*. 804 F.3d 202 (2d Cir. 2015).

79 Andres Guadamuz, 'Google and Book Publishers Settle' (*WIPO Magazine*, July 2009) <https://www.wipo.int/wipo_magazine/en/2009/04/article_0004.html> accessed 25 November 2023.

80 ibid.

81 ibid.

82 See: Siva Vaidhyanathan, 'The Googlization of Everything and the Future of Copyright' (2007) 40(3) University of California Davis Law Review 1207-231; and Alison Flood, 'Ursula Le Guin leads revolt against Google digital book settlement' *The Guardian* (London, 22 January 2010) <https://www.theguardian.com/books/2010/jan/22/ursula-le-guin-revolt-google-digital> accessed 25 November 2023.

83 *Authors Guild v Google, Inc*. 770 F. Supp. 2d 666 - Dist. Court, SD New York 2011.

84 *Authors Guild v Google* (2015), at 214.

85 ibid at 223.

86 Michael W Carroll, 'Copyright and the Progress of Science: Why Text and Data Mining Is Lawful' (2019) 53 UC Davis Law Review 893; and Matthew Sag, 'Copyright and Copy-Reliant Technology' (2009) 103 Northwestern University Law Review 1607; Jessica Gillotte, 'Copyright Infringement in AI-Generated Artworks' (2020) 53 UC Davis Law Review 2657; Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning' (2019) 66 Journal of the Copyright Society of the USA 291; and Van Lindberg, 'Building and Using Generative Models Under US Copyright Law' (2023) 18 Rutgers Business Law Review 1.

87 Yvette J Liebesman and Julie C Young, 'Litigating Against the Artificially Intelligent Infringer' (2020) 14 FIU Law Review 259 <https://ecollections.law.fiu.edu/lawreview/vol14/iss2/8> accessed 25 November 2023.

88 Benjamin LW Sobel, 'Artificial Intelligence's Fair Use Crisis' (2017) 41 Columbia Journal of Law & the Arts 45, 80.

89 See above n 1.

90 Countries include Japan, Singapore, Ecuador, the UK, and the EU, see: Sean M Fiil-Flynn and others, 'Legal Reform to Enhance Global Text and Data Mining Research' (2022) 378 Science 951.

91 Intellectual Property Office, 'Digital Opportunity: A Review of Intellectual Property and Growth' (2011) <https://assets.publishing.service.gov.uk/media/5a796832ed915d07d35b53cd/ipreview-finalreport.pdf> accessed 25 November 2023.

92 ibid at 45-47.

93 Implemented in the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, which introduced s 29A to the CDPA.

94 s 29A 1(a) CDPA.

95 s 29A 1(b) CDPA.

96 Such as that contained in s 29 CDPA.

dataset. There are dozens of research projects that collect music, data or images, and all those collections would fall under this exception.[97] If the exception applies to the gathering phase, it also makes sense that it covers the training phase too, as these technical stages are not contemplated in the legislation. What matters is that this operation could involve the copying of data to generate a model.

There are a couple of outstanding questions with regards to the use of this exception. Firstly, there is no standard definition of what research is: many tech companies have large research bodies, but it would be fair to assume that these would be excluded, as Sec. 29A specifically refers to 'non-commercial research'. Interestingly, there has never been a judicial definition of what constitutes research,[98] but there is agreement that commercial research would not fall into this category.[99]

Secondly, there is the question of whether a dataset collected for research can be reused by a commercial entity. Some datasets are available to the public, especially datasets that may have to be shared due to funding obligations. There is nothing in the law that specifies that subsequent uses of the dataset must be also non-commercial, opening the door to what some have named data-laundering, or academic-washing.[100] In these situations a private company uses a research dataset to produce commercial works and this is actually the case with WebVid, which is being used by Meta as one of the sources for training[101] their video generation models called Make-A-Video.[102] There is currently no answer to this second question in the UK, and this appears to be a loophole in the law which goes against the spirit of having a research exception in the first place.

However, this question may soon become moot. In 2021, the UK Intellectual Property Office conducted a consultation on several aspects of IP law and artificial intelligence in the UK, and the government responded in 2022.[103] One of the questions in the consultation was precisely about the existing TDM exception, asking respondents to choose from various options. These were to do nothing; improve the licensing environment to include TDM; expand the TDM exception to include commercial uses; expand the exception for all purposes, with an opt-out for authors; or expand the exception for all purposes, not allowing for an opt-out.

The responses were varied, but in the end the UK Intellectual Property Office (UKIPO) decided to opt for the last option, which would make it the most open TDM exception in the world.[104] One of the objectives for the adoption of an expanded exception was to put the UK in a relative competitive advantage when it comes to artificial intelligence research. The UK government saw the adoption of an AI-friendly policy as helping to generate income and attract investment, making 'the most of the greater flexibilities following Brexit'.[105]

However, this reform has received considerable criticism, the most comprehensive of which came in a report issued by the House of Lords Communications and Digital Committee,[106] which called the UKIPO proposal 'misguided' and asked the UKIPO to halt the reform and open further consultations.[107]

While the reforms appeared to have been halted at some point,[108] the UK government announced that it would move forward with a reform to TDM exception as 'there remains a lack of regulatory clarity as to the direction of those reforms'.[109] While the text of any change to the existing exception was not published, the content of the recommendations by Sir Patrick Vallance contained language that pointed towards a pro-investment and pro-innovation reform that would 'prioritise practical solutions to the barriers faced by AI firms in accessing copyright and database materials'.[110] At the time of writing, the Government has published a response to the Vallance Report which concludes that the UKIPO will draft a code of conduct for AI firms, and this will be drafted in consultation with 'a group of AI firms and rights holders to identify barriers faced by users of data mining techniques when accessing copyright material'.[111]

While a code of conduct could be useful, it is clear that a legislative reform would be better, and is sorely needed, at the very least to close the data laundering loophole, but

97 Another example is a dataset of heavy metal music, which also contains videos and interviews, see: Jan Herbst and Mark Mynett, 'Heaviness in Metal Music Production' (2022) <https://doi.org/10.34696/9s05-wv03> accessed 25 November 2023.

98 Abbe Brown and others, *Contemporary Intellectual Property: Law and Policy* (5th edn, OUP 2019) 177.

99 An example of a dataset that would not be included because of this would be HD-VILA-100M, a high-resolution video dataset collected from YouTube by researchers from Microsoft Research. See: Hongwei Xue and others, 'Advancing High-Resolution Video-Language Representation With Large-Scale Video Transcriptions' (2022) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 5036-5045 <https://perma.cc/KDY4-H3UM> accessed 25 November 2023.

100 Andy Baio, 'AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability' (*Waxy Blog*, 30 September 2022) <https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/> accessed 25 November 2023.

101 Uriel Singer and others, 'Make-A-Video: Text-to-Video Generation without Text-Video Data' (*arXiv*, 29 September 2022) <http://arxiv.org/abs/2209.14792> accessed 25 November 2023.

102 <https://makeavideo.studio/> accessed 25 November 2023.

103 UK Intellectual Property Office, 'Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation' (2022) <https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents> accessed 25 November 2023.

104 ibid.

105 ibid.

106 House of Lords Communications and Digital Committee, At risk: our creative future (2023), 2nd Report of Session 2022-23 HL Paper 125 <https://committees.parliament.uk/committee/170/communications-and-digital-committee/news/175423/dont-let-complacency-jeopardise-the-creative-industries/> accessed 25 November 2023.

107 ibid, at 34-35.

108 Hansard, HC deb 1 February 2023 vol 727 Col 152WH. See also: Rory O'Neill, 'UK government bins UKIPO's flagship AI reforms' (*Managing IP*, 3 February 2023) <https://www.managingip.com/article/2b8dy58efmhhbvsmaxvk0/uk-government-bins-ukipos-flagship-ai-reforms> accessed 25 November 2023.

109 HM Government, 'Vallance Review: Pro-innovation Regulation of Technologies Review Digital Technologies' (2023) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1142883/Pro-innovation_Regulation_of_Technologies_Review_-_Digital_Technologies_report.pdf> accessed 25 November 2023.

110 ibid.

111 HM Government, 'HM Government Response to Sir Patrick Vallance's Pro-Innovation Regulation of Technologies Review' (March 2023) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1142798/HMG_response_to_SPV_Digital_Tech_final.pdf> accessed 25 November 2023.

also because as things stand there is a great uncertainty in an area that is starting to generate litigation. It is disappointing that the push towards reform has been stopped, at least for now.

One of the reasons for the UK's regulatory action in this area has been precisely to respond to developments in Europe.[112] The EU adopted its own TDM exceptions in 2019 as part of the Digital Single Market (DSM) Directive.[113] The DSM Directive deals with a wide range of digital copyright issues, and it implements two different exceptions with regards to data mining that are directly relevant to AI inputs.

In Art. 3, the DSM Directive sets out a new exception for copyright for 'reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.'[114] This is in many ways like the UK exception, perhaps with more precise language. Article 2 of the Directive also usefully defines a research organisation as a university, its libraries, or any other research entity whose 'primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research'.[115] This research is carried out on a non-profit basis (including reinvesting all profits into scientific research), or pursuant of a public interest mission recognised by the Member State.

Article 4 provides the biggest change, as it extends this exception to commercial organisations for reproduction and extraction for the purpose of data mining, if they have lawful access to the work, and the work 'has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.'[116]

As the DSM Directive has only been in effect a relatively short time, there has not been time for it to be tested in court,[117] so the precise interpretation of its provisions is still open to analysis. Here the recitals offer a very important glimpse into the possible limits and interpretation. Just as with the UK norm, the stated purpose of these exceptions is to facilitate scientific research. While many uses may be already fair or fall under existing exceptions such as temporary copies,[118] there is still a need for an exception because some uses could be infringing, and therefore the EU's 'competitive position as a research area will suffer, unless steps are taken to address the legal uncertainty concerning text and data mining.'[119] The commercial exception is being implemented because several research institutions have commercial arms, as well as public-private partnerships and collaborations with the private sector, which would benefit from a wider exception that allows for such uses.[120] The opt-out option is given to rightsholders to provide balance between their interests and those of the user, as it is clear that the Directive is attempting to pass the exceptions in compliance with the three-step test.[121]

There are two interesting and relevant aspects covered in the recitals with regards to Art. 3. The first is that rightsholders can apply security measures to ensure the integrity of their systems and databases, though these should be proportional and not excessive.[122] The second is that the application of the exceptions does not need to be accompanied by compensation for rightsholders.[123]

Generally, all the above should work to allow some machine learning operations to take place legally, but there will be some room for interpretation depending on the particulars of each situation. Some national legislation has been adding more detail to Arts. 3 and 4, which could further help to clarify any doubts. For example, in their transposition of the DSM Directive,[124] France introduces a version of both exceptions, with the added detail that the opposition to Art. 4 can be done 'by means of machine-readable processes for the content made available to the public online'.[125] This opt-out does not need to be justified by the rightsholder.[126] Germany has also included a similar provision allowing machine-readable opt-outs.[127] Interestingly, the German transposition of Art. 3 contains the exception for scientific research, and it also incorporates the definition of research organisation contained in Art. 2 DSM Directive.[128] This section specifically excludes organisations that are collaborating with private enterprises because these would fall under the commercial exception and the organisation would therefore have to respond to opt-outs in accordance to Art. 4 (§ 44b Copyright Act, UrhG).

One thing that is certain is that at least until now there has been considerable interest in policymaking circles to have some form of exception for data mining, and this covers several acts of training data. It will be interesting

112 The Directive will not be adopted in the UK due to Brexit, so it will not be covered in a lot of depth here. For more on the Directive, see: Giancario Frosio, Christophe Geiger and Oleksandr Bulayenko, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU' in Concepción Saiz García and Raquel Evangelio Llorca (eds), *Propiedad intelectual y mercado único digital europeo* (Tirant lo blanch 2019) 27-71; and Christophe Geiger, 'The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive' in Martin Senftleben and others (eds), *Intellectual Property and Sports, Essays in Honour of P. Bernt Hugenholtz* (Kluwer Law International 2021) 383-94.

113 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L130/92-125.

114 ibid art 3.

115 ibid art 2.

116 ibid art 4.

117 At the time of writing there is an ongoing case in Germany against LAION which may explore this issue. See: Andres Guadamuz, 'Photographer sues LAION for copyright infringement' (*Technollama Blog*, 5 May 2023) <https://wp.me/pwh3D-3ur> accessed 25 November 2023.

118 ibid Recital 9.

119 ibid Recital 10.

120 ibid Recital 11.

121 ibid Recital 6. For more on the 3-step test, see: Martin RF Senftleben, *Copyright, Limitations, and the Three-Step Test: An Analysis of the Three-Step Test in International and EC Copyright Law* (Kluwer Law International 2004).

122 Recital 16.

123 Recital 17.

124 In Ordonnance n° 2021-1518 du 24 novembre 2021, and Décret n° 2022-928 du 23 juin 2022.

125 Contained in art 122-5-3 of the Code de la propriété intellectuelle. It can also be done by a letter.

126 art 122-32 Code de la propriété intellectuelle.

127 § 44b Urheberrechtsgesetz (UrhG).

128 § 60d UrhG.

to chart any changes to these AI-friendly policies, because there could be a backlash to allowing the indiscriminate training of AI using copyright works. The existing exceptions were drafted with very specific types of data mining in place, with the fight against disease and the development of new medicines being cited repeatedly to justify these exceptions. Now, though, policymakers will have to contend with angry rightsholders that see their works used in machine learning without equitable remuneration, and it will be interesting to see how this possible conflict plays out.

IV. Copyright infringement in outputs

While the input picture is complicated, the output question may be slightly easier to analyse. It is tempting to stop at the input stage because if there is infringement there, rightsholders would not need to try to prove infringement in the outputs. However, it could be that the training falls under fair dealing, either because of a temporary copy, or because of the TDM exception. So, it may be useful to try to see if there is infringement in an output. Assuming the training phase has passed the infringement test (which is a big assumption) could there be infringement in the outputs?

For there to be copyright infringement three requirements need to be met: (i) the infringer undertook one of the exclusive rights of the author without authorisation;[129] (ii) there is a causal connection between both works; and (iii) the entirety of the work, or a substantial part of it, has been copied.[130]

1. The exclusive rights of the author

The first requirement is self-explanatory, but it may require further analysis to try to picture what an infringement of an AI output would look like. For now we will only look at the rights of reproduction and adaptation, although there could be other infringements in the input phase.[131]

With regards to copying, this could perfectly take place with regards to inputs, namely in making copies to use in the training of a model, but can we talk about copying with an output? This could be difficult to prove, as we may have to look at the technology in detail to see if an output constitutes a copy. Strictly speaking, a copy of a work is made when there is a reproduction in any form of the whole or part of a work.[132] So a sculpture of a photograph,[133] a photocopy of a book, or an mp3 file of a vinyl record would all be copies.[134] The main question here may not be one of exact reproduction, but rather a conversion of a work into something that is different to the original work, something we will deal with while covering the third requirement.

When works are different, we may be faced with an adaptation (in other jurisdictions this is called a derivative work).[135] In the UK, an adaptation has a specific meaning, namely the exclusive right of the author to allow for the translation of a literary work, or to make a non-dramatic work into a dramatic one (and vice versa), or to turn the literary work into a work with images,[136] amongst other actions.[137] For the sake of brevity, we will treat both reproductions and adaptations interchangeably here, although a reproduction could actually be an adaptation under some circumstances.

The question of reproduction may end up being a technical one. The emergence of AI public tools has prompted a debate on whether these tools can reproduce a work in the training data in an output. There appears to be some evidence that on some occasions this can take place, but it is relatively rare, and sometimes researchers must purposefully set out to achieve it. A study on Github's Copilot found that code recitation (in legal terms, a reproduction) took place in 0.1% of the times.[138] To put this figure in context, the study looked for code that is reproduced exactly in 60 words at least, and found that out of 453,780 code suggestions, only 473 matched some of the training code at that level. This is a small number, but the most important aspect of the analysis was the nature of the code that was being replicated. It tended to be common elements of code, mostly opening text. Copilot was more likely to suggest code from somewhere if there was not a lot of input, and as more input was offered, the less likely it was to offer matching code.[139] There have been other claims of code recitation in Copilot,[140] but nothing specific. In fact, the ongoing class action lawsuit against Github and OpenAI has not produced any specific examples of code recitation.[141]

Another study found some reproduction in Copilot, but not many examples of verbatim replication.[142] The same happens with other language models such as GPT-2 and GPT-3,[143] where there is evidence of what the researchers call memorisation of common data in the dataset. The best example of this sort of memorisation in LLMs is in

129 Namely to do any of the following without authorisation by the copyright holder: copy, issue copies, perform, rent, lend, communicate to the public, or adapt the work. See s16 (1) CDPA.

130 Lionel Bently and others, *Intellectual Property Law* (5th edn, OUP 2018) 194.

131 As explained above, communication to the public could potentially be infringed in inputs.

132 s 17 CDPA.

133 Specifically referred to in s 17(3) CDPA.

134 See Caterina Sganga, 'The Right of Reproduction' in Eleonora Rosati, *The Routledge Handbook of EU Copyright Law* (Routledge 2021) 126.

135 See: Pamela Samuelson, 'The Quest for a Sound Conception of Copyright's Derivative Work Right' (2012) 101 Georgetown Law Journal 1505.

136 s 21 CDPA.

137 There are different rules for computers programs, databases, and musical works.

138 Albert Ziegler, 'GitHub Copilot research recitation' (*GitHub Blog*, 30 June 2021) <https://github.blog/2021-06-30-github-copilot-research-recitation/> accessed 25 November 2023.

139 ibid.

140 See for example: <https://twitter.com/DocSparse/status/1581461734665367554> accessed 25 November 2023.

141 See the defendant's response asking for dismissal <https://fingfx.thomsonreuters.com/gfx/legaldocs/xmvjklzrypr/MICROSOFT%20OPENAI%20LAWSUIT%20openaimtd.pdf> accessed 25 November 2023.

142 Daphne Ippolito and others, 'Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy' (*arXiv*, 31 October 2022) <http://arxiv.org/abs/2210.17546> accessed 25 November 2023.

143 ibid.

a study[144] that looked at whether a model could recite the next line of a book and found that popular books that were probably prevalent in the input data could display such memorisation. Another recent study found that language models could be tricked into spewing random memorised text.[145]

There have also been examples of memorisation of inputs in a couple of studies conducted with image AI tools. Specifically with regards to machine learning, any possible reproduction of images found in the training data can happen due to what is known as overfitting.[146] This is understood as the memorisation of individual items in the training data, particularly those which are repeated often. With regards to images, a study by Somepalli et al[147] found some memorisation in image datasets, and was therefore able to reproduce some items found in the input. They defined replication as an image that contains an object 'that appears identically in a training image, neglecting minor variations in appearance that could result from data augmentation'.[148] The study looked at various datasets and models, and found very low replication rates with the ImageNet dataset.[149] There was more significant replication picking from the LAION Aesthetics dataset, which consists of 12 million images. As has been explained before, LAION consists of a link to an image, as well as the ALT text used to describe the image. The researchers took some of this text, ran it on a Stable Diffusion model on a random sample of 9,000 prompts, and found 170 images with very high similarity to the original. By running several experiments on that dataset, the paper found that 1.88% of random generations had a high similarity score with the training material, which is still a relatively high incidence. The reason for some of the high similarity results was, unsurprisingly, the prevalence of popular and repeated images such as famous paintings.[150] In the discussion of the limitation of the study, the researchers comment that they used a small dataset, while LAION consists of billions of images, so their sample was only 0.6% of the total training data.[151]

Another study by Carlini et al.[152] was also able to reproduce some images contained in the training data. Although this study appears to be more concerned with finding possible privacy violations,[153] the results are still interesting. The researchers used a model from Stable Diffusion trained on 160 million images, choosing 350,000 of the most replicated samples in the dataset.[154] For each image, they generated 500 samples, producing a total of 175 million results. The replication rate was low, however, producing a total of 50 memorised images.[155]

So, it would be fair to say that there is certainly some form of memorisation across various models, but the rate tends to be low. It is also important to point out that some of these studies are setting out to find replication as an attack vector to generate measures to reduce the replication. While this does not invalidate the results, it is relevant to point out that memorisation, while possible, is relatively rare and can potentially be ameliorated.

It is also relevant to point out that memorisation as described in these papers may not be the same as a reproduction; after all, the term 'memorisation' is not legal, and it is not an exclusive right of the author (humans can also memorise material without infringing). So, the legal question is simple: has a work in the input been copied? From a legal perspective, what is important is that the generation of copied images from the dataset could fall under copyright infringement, but it seems unlikely that this replication is at any large scale, at least based on the research highlighted above. One could argue that the owners of a replicated image could sue for copyright infringement, but the rates described in the papers would make it difficult for any random rightsholder to find some replication of their work.

Another thing to consider here is that while exact reproduction is rare, machine learning models are good at generating styles of creators and artists that are popular in the training data. So, asking ChatGPT for a poem in the style of Richard Brautigan or Pablo Neruda will produce results that reasonably resemble those of the writers,[156] as will asking for lyrics in the style of Shakira.[157] The same goes for famous paintings, where image generators can reproduce works such as the Mona Lisa and Starry Night reasonably well because they can be found everywhere in the training data (Image 3).

So, there can undoubtedly be some form of reproduction in outputs, but these tend to be inexact copies, and in the case of image models they can be dependent on how common they are in the training data, and would therefore be more about the analysis given to inputs.

A final note on reproduction is that some technologies may make the need for a legal analysis of what is copying in AI necessary sooner rather than later. Stable Diffusion has developed a model called ControlNet,[158] which takes a source image, even the sketchiest, and produces a viable output from it. So, you can use a child's drawing, or even a stick figure, and the system will provide a full version of it (Image 4). Whether this would be considered reproduction, or even a translation, will have to be determined in the future.

144 Kent K Chang and others, 'Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4' (*arXiv*, 28 April 2023) <http://arxiv.org/abs/2305.00118> accessed 25 November 2023.

145 Milad Nasr and others, 'Scalable Extraction of Training Data from (Production) Language Models' (*arXiv*, 28 November 2023) <http://arxiv.org/abs/2311.17035> accessed 25 November 2023.

146 Devansh Arpit and others, 'A Closer Look at Memorization in Deep Networks' (*arXiv*, 1 July 2017) <http://arxiv.org/abs/1706.05394> accessed 25 November 2023.

147 Gowthami Somepalli and others, 'Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models' (*arXiv*, 12 December 2022) <http://arxiv.org/abs/2212.03860> accessed 25 November 2023.

148 ibid 3.

149 ibid 7.

150 ibid 9.

151 ibid 11.

152 Nicholas Carlini and others, 'Extracting Training Data from Diffusion Models' (*arXiv*, 30 January 2023) <http://arxiv.org/abs/2301.13188> accessed 25 November 2023.

153 ibid 2.

154 ibid 5.

155 ibid 7.

156 <https://twitter.com/gikii/status/1627091146462863362> accessed 25 November 2023.

157 <https://twitter.com/gikii/status/1627093439958646784> accessed 25 November 2023.

158 Lvmin Zhang, Anyi Rao and Maneesh Agrawala, 'Adding Conditional Control to Text-to-Image Diffusion Models' (*arXiv*, 10 February 2023) <http://arxiv.org/abs/2302.05543> accessed 25 November 2023.

Image 3. *Starry Night and the Mona Lisa, generated with Midjourney*

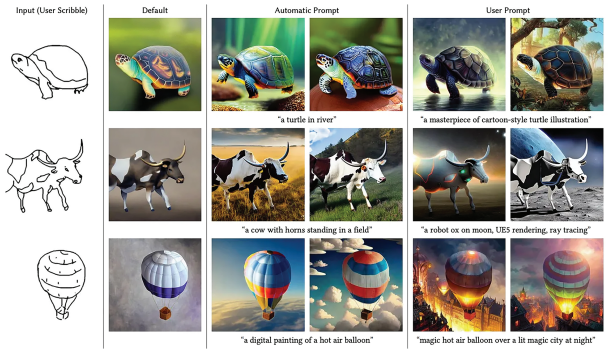

Image 4. *ControlNet in action*[159]

2. Causal connection

The second requirement for infringement in the UK is that there must be a causal connection between the original work and the alleged infringing copy. This is to avoid cases of independent creation where one work resembles another by coincidence, or because both authors were inspired by similar works.

A large amount of copyright litigation rests on finding these connections. For example, in the famous case of *Francis Day v Bron*,[160] two songs were found to be similar, but the court could not find a causal connection between the author and the alleged infringer; claimants must not only prove similarity, but that this similarity was due to an act of copying. Another example can be found in *Mitchell v BBC*,[161] where two sets of children's TV characters were found to be similar, but the similarity arose mainly because the creators had been inspired by similar sources.[162] Birss J commented that any resemblance was due to 'artists have worked in the same field of children's character design and no doubt have been influenced by the many of the same common elements'.[163]

More recently in *Sheeran v Chokri*,[164] the singer and songwriter Ed Sheeran was sued for copyright infringement by Sami Chokri over the song 'The Shape of You'. Chokri alleged that Sheeran had copied parts of his melody 'Oh Why', and while there were definitely some similarities between both compositions, Zacaroli J found that Sheeran had used that musical pattern before; furthermore Chokri's composition had only played in the radio once, and there was no evidence that Sheeran had ever listened to it.

Proving a causal connection between artificial intelligence inputs may be easier than the arduous collection of evidence in some of the cases described above. In some instances, the training datasets are public, so it is easier to scan through them to try and find if a work is in the training data. A causal connection could be established if there is some sort of replication in the output. However, not all datasets are public, with some companies – such as OpenAI and Google – keeping some of their datasets closed, making it difficult to prove a connection in some cases.

Moreover, there could be other complications depending on the type of work. In the case of computer code, for example, some form of replication in the outputs could originate from there being a popular solution to a common problem. That could make pinpointing an original source code replicated in an output difficult considering the idea/expression dichotomy.[165] In the case of Copilot, research found that a lot of code recitation took place from sources that offered a common solution that was part of widely shared code, such as code used in programming classes, or dealing with specific hardware problems in toys and robotics.[166]

With regards to images, sometimes it is evident that a model has been trained with a specific dataset. Stable Diffusion can sometimes reproduce watermarks from image repositories like Getty (Image 5), but the source of the actual image is not in the Getty collection, or

Image 5: *Stable Diffusion v 1.5 output showing the Getty Images logo*[167]

159 ibid.
160 *Francis Day & Hunter v Bron* [1963] Ch 587.
161 *Mitchell v BBC* [2011] EWPCC 42.
162 ibid 94.
163 ibid 145.
164 *Sheeran & Ors v Chokri & Ors* [2022] EWHC 827 (Ch).

165 This can be found in cases such as *Designers Guild Ltd v Russell Williams (Textiles) Ltd* [2000] 1 W.L.R. 2416. See also its role in software in: Noam Shemtov, 'The Idea–Expression Dichotomy and Its Role in Software-Related Disputes' in Noam Shemtov (ed), *Beyond the Code: Protection of Non-Textual Features of Software* (OUP 2017).
166 Ziegler (n 138).
167 It is worth pointing out that later versions of Stable Diffusion do not reproduce logos or watermarks.

at least a reverse search did not produce any hits. So a model 'learns' that some images have watermarks and will place it sometimes, if prompted to do so, but the output itself does not come from Getty Images: it is a generative creation built by the model taking data points from latent space, which it then adds to the logo upon request.

Another potential problem in establishing causal connection is the existence of large amounts of material in the training data that resembles a work but does not originate from the owner of that work. This is the case of so-called fan fiction, or fan art.[168] One of the challenges for living creators, but also for others whose work may still be under copyright such as Picasso or Jean-Michel Basquiat, is that an output may be the result of training on hundreds of human imitators that can be found all over the web.[169] So let us consider an artist who has successfully removed their works from a dataset by taking advantage of the opt-outs in Art. 4 of the DSM Directive, but their work is still reproduced by an AI tool by producing a recognisable style attributed to that artist. This could be because the training dataset contains works from human imitators.

There may be a technical way to find a causal connection between a work in the input, but at least at the time of writing there is no tool that can perform this action. Research is underway[170] to try to find methods for looking for similarity in works, but this is still in its early stages.[171] The takeaway for the purpose of this paper is that at present it's not possible to establish direct correlation between inputs and outputs, but the existence of such technology could provide methods to avoid litigation by proving the absence of a connection between two works.

3. Substantial copying

The final requirement is that the whole of the work, or a substantial part of it, has been copied.

The main issue here will be that there must be something in the output that has been copied or transformed to begin with. Exact replication is likely to be rare given the vast amounts of training datapoints mentioned above, and the relatively small number of verbatim copies found in the literature, even when setting out to try to obtain a replica. So, most of the potentially infringing outputs would be partial or inexact copies, and the legal question becomes one of similarity between the input and the output.

The main authority in the UK analysing substantial infringement is the case of *Designer's Guild v Russell Williams*,[172] where the claimant had created a textile flower design and brought an infringement suit against the defendant alleging that copying had taken place. The original flower design had impressionistic red lines and coloured flowers scattered across, while the alleged copy also had lines and flowers, though in a different arrangement. A causal connection had been established, so the question rested on how substantial the copying was. The trial judge found infringement, the appeal made it to the House of Lords, where it was held that there was indeed substantial infringement. What is relevant to the present analysis is that when considering whether copying had been substantial, the analysis should be qualitative and not quantitative, so if a small part of a work had been copied, but that part was very important to the whole, then there would be infringement.[173] Another case where this is made clear is *England And Wales Cricket Board v Tixdaq*, where eight seconds of video footage from a cricket match were found to be substantial copying if they showed an important event, such as something interesting happening.[174] The test of what is substantial has become tied with what gives the work originality in the first place;[175] a substantial part is that which makes the work worthy of protection in the first place, and that is a qualitative test.

European cases have also dealt with the substantiality requirement. *Infopaq I*[176] addressed issues of copyright law related to substantial copying and temporary acts of reproduction. Infopaq, a media monitoring company, sent customers summaries of newspaper articles selected by a data capture process, which involved scanning articles, processing them to identify relevant words, and printing extracts. The Danish Publishers Association (DDF) contested this, claiming copyright infringement. The CJEU ruled that an 11-word extract of a protected work is a reproduction under Art. 2 of Directive 2001/29/EC if it expresses the author's intellectual creation.[177]

It could be argued that the qualitative test means that for there to be infringement, the amount of the original work that is present in the output can be relatively small, if it is an important part of the work. But there must be something in the output that resembles the input in the first place, and this is not easy to ascertain in AI outputs. With millions, and sometimes billions of input works, there may not be a trace of anything in an output. To claim otherwise would reduce copyright to something akin to saying that every artist that has been inspired by other people's work would also be infringing, which leads to the absurd conclusion that potentially every work in the world is infringing everything else. We have established that training a model does not generate a repository of exact copies that become mixed into some form

168 For a look at this phenomenon and copyright, see: Aaron Schwabach, *Fan Fiction and Copyright: Outsider Works and Intellectual Property Protection* (Routledge 2016).

169 Art repository websites such as Bēhance, DeviantArt, and ArtStation, are filled with imitations of living and dead artists, as well as fan art of popular culture characters.

170 Nikhil Vyas, Sham Kakade and Boaz Barak, 'Provable Copyright Protection for Generative Models' (*arXiv*, 21 February 2023) <http://arxiv.org/abs/2302.10870> accessed 25 November 2023.

171 Boaz Barak, 'Provable Copyright Protection for Generative Models' (*Windows on a Theory Blog*, 21 February 2023) <https://windowsontheory.org/2023/02/21/provable-copyright-protection-for-generative-models/> accessed 25 November 2023.

172 *Designer Guild Ltd v Russell Williams (Textiles) Ltd* [2001] FSR 113.

173 ibid.

174 *England and Wales Cricket Board Ltd & Anor v Tixdaq Ltd & Anor* [2016] EWHC 575.

175 See also *The Newspaper Licensing Agency v Marks and Spencer* [2003] 1 AC 551, at 559.

176 *Infopaq International A/S v Danske Dagblades Forening* (n 59).

177 See also Case C-476/17 *Pelham v Hütter* ECLI:EU:C:2019:624, [2019] ECDR 26.

Image 6. *A cat by Van Gogh in Brighton, and a llama by Banksy, made with Midjourney*

of collage in an output, and what happens is more like memorisation of what something looks like by the accumulation of data points that get stored in a latent space. True, there could be some works that get memorised through overfitting, but this is a rare occurrence. For the most part a system like ChatGPT is acting on a statistical likelihood of words following one another in a text, so it is not a verbatim copy of a work.[178] And even in the cases of memorisation, this may not be infringement, as memorisation is not an exclusive right of the author.

AI tools may be able to replicate styles, such as writing or artistic styles, and write some text in the style of an author or produce an image in the style of an artist, which would not have copied the original in any way. So, are styles protected by copyright? Roughly speaking no. A style is more of an idea, and copyright does not protect an idea, only the expression of that idea,[179] because protecting an idea would potentially lead to monopolies.[180]

While it is possible to produce an image in an artist style with DALL-E or Midjourney, that image may not be infringing. One can ask for a painting of 'a cat by Van Gogh in Brighton' and get a very good image in Van Gogh's unmistakable style, but he never went to Brighton, and never painted a cat, so were he alive it would be difficult to imagine a successful infringement suit based on the style. One can also prompt for 'a llama by Banksy', and similarly generate something that could have been made by Banksy even though that artist has never generated an image of a llama (Image 6).

Here is where the idea/expression dichotomy will be at the forefront of the legal analysis of the similarity between two works. Popular artists and writers will be reproduced in the models repeatedly, so the question may rest on how similar the works are, i.e. is there substantial reproduction? One important test case here is *Temple Island Collections v New English Teas*.[181] The case involves a black and white image of the UK Parliament building, and a bright red bus travelling across Westminster Bridge. The claimant owned the photograph which was used in London souvenirs, and the defendant was a tea company that created a similar picture for a publicity campaign. A causal connection had been established as the defendants had wanted to licence the claimant's picture, and when that negotiation led nowhere, they produced their own version. The images were undeniably similar, but was there substantial similarity? Birss QC followed the qualitative test from *Designer's Guild* and found that while not every important part of the composition in the original picture had been reproduced, both images had substantial elements that amounted to infringement.[182] But what may be important for future similarity tests regarding images is that there must be common elements between the two pictures: Birss QC found similarity in the composition, the colouring, and also several objects in both pictures. So, not just a passing similarity of styles will amount to copyright infringement – those visual elements must have 'visual significance'.[183]

One could argue that *Temple Island* treads dangerously close to protecting an idea,[184] but at least it gives us some indication that for visual works the substantial similarity must replicate significant visual elements, and not just a general style.

Interestingly, copyright holders may have a better claim of substantiality with outputs that reproduce works that are under copyright protection such as artwork, characters, and popular media. AI tools can do a good job of reproducing some of these because the sources are repeated in the training data. So, you can ask an image tool such as Midjourney or Stable Diffusion's Dreamstudio to reproduce a famous character, and it will often manage to do a decent job of it (Image 7).

Leaving aside the important issue of moral rights for now,[185] it may be easier to see how any of the above could potentially be found to be infringing, particularly with regards to character copyright,[186] but there could also be claims using other types of IP, such as trademarks.[187] Enforcement of these possible infringing copies may depend entirely on the person making them, and a commercial entity would probably find itself at the end of successful litigation if it used such images without authorisation. But it is difficult to see how an individual user would be subject to litigation because the internet is already filled with similar infringing uses that

178 Stephen Wolfram, 'What Is ChatGPT Doing … and Why Does It Work?' (*Stephen Wolfram Writing*, 14 February 2023) <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/?fbclid=IwAR13pTPLTEPFem8g71AnUZ-b7xRHAp69b027lbPt8DVdka1uTX07MhDkvyqU> accessed 25 November 2023.

179 *Designer Guild Ltd* (n 172). See also: Deming Liu, 'Reflections on the Idea/Expression Dichotomy in English Copyright Law' (2017) 1 Journal of Business Law 71.

180 *Jones v London Borough of Tower Hamlets* [2001] RPC (14) 379.

181 *Temple Island Collections Ltd v New English Teas Ltd and Another (No. 2)* [2012] EWPCC 1.

182 ibid 63.

183 ibid 34.

184 Andreas Rahmatian, 'Temple Island Collections v New English Teas: An Incorrect Decision Based on the Right Law?' (2012) 34 European Intellectual Property Review 796.

185 For more on that subject, see: Rita Matulionyte, 'Can AI Infringe Moral Rights of Authors and Should We Do Anything About It: An Australian Perspective' (2022) <https://ssrn.com/abstract=4016001> accessed 25 November 2023.

186 For more on this, see: Zahr K Said, 'Fixing Copyright in Characters: Literary Perspectives on a Legal Problem' (2013) 35 Cardozo Law Review 769; Katerina Sharkova, 'The author, the fan and the in-between: in search of a copyright regime for the everyday creative' (2018) 40 EIPR 784-96; and Samuel J Coe, 'The Story of a Character: Establishing the Limits of Independent Copyright Protection for Literary Characters' (2011) 86 Chicago-Kent Law Review 1305.

187 For more on that, see: Ilanah Simon Fhima, *Trade Mark Dilution in Europe and the United States* (OUP 2011).

Image 7. *Superman shopping, and Pikachu at a bar, made with Midjourney*

are generated by humans which do not get enforced for a variety of reasons (but mainly because its impractical, expensive, and it is not a good practice to sue your fans). However, large-scale infringement of this nature could eventually be the subject of a lawsuit, allowing us to explore interesting issues beyond substantiality to the possible use of exceptions.

4. Exceptions for outputs

The three exceptions that could be used for outputs are caricature, parody, and pastiche.[188] There has been some debate about whether these are the same category, or three different exceptions worthy of separate definitions. In in discussion leading the landmark CJEU case of *Deckmyn v Vandersteen*,[189] AG Cruz Villalón believed that the concepts 'are not in competition with one another'[190] *Deckmyn* dealt with parody separately, so for the purpose of this article we will do so as well.[191]

With regards to parody, we have had a legal definition of what constitutes it from the CJEU:

> '… the essential characteristics of parody, are, first, to evoke an existing work, while being noticeably different from it, and secondly, to constitute an expression of humour or mockery.'[192]

Caricature itself has not been defined by the courts. The Oxford English Dictionary defines it as a 'portrait or other artistic representation, in which the characteristic features of the original are exaggerated with ludicrous effect.'[193] Griffith usefully points out that the dictionary definition appears to closely match that of parody as found in *Deckmyn*,[194] so both can receive similar analysis.

From the above there is certainly a degree of intention needed to produce a parody or caricature; there must be an element of mockery or exaggeration; and it must both clearly refer to a work and be recognisably different from it. It is possible that one could set out to produce a work of parody or caricature with AI, and the examples in Image 7 could potentially be considered a parody, especially if one drafts a prompt asking for a parody specifically.[195] Language models like ChatGPT can also produce parodies of works. For example, I asked for a parody of the Richard Brautigan's poem 'All watched over by machines of loving grace', and the program produced an amusing poem entitled 'All watched over by machines of mugging face'.[196] There is no reason to believe that such parodies generated with AI tools are not covered by the parody exception if they resemble the source closely.

The pastiche exception is of special interest for the subject of transformative uses.[197] The issue of pastiche is generally less explored in the literature, and has rarely been tested in the courts,[198] so the concept's definition is still up for debate. Hudson says that it is often seen as 'laudatory and non-critical imitation, such as creating a new work in the style of another artist or genre, and making a new work from a compilation or assembly of pre-existing works'.[199] Alternatively, Flaherty proposes a definition of pastiche using analogous decisions in other fair dealing cases which explains that it is a work that evokes another, but 'must be noticeably different from that first work, displaying original thought such as to clearly represent a new work',[200] and must contain a laudatory comment on the author of the cited work.

A case in Germany could bring more clarity to the subject.[201] It involves an English artist, Daniel Conway, who made a digital image called 'Scorched Earth'. A German artist, Martin Eder, found the image in a cheap paint-by-numbers book on Amazon for $12, and combined it with other images, including an 1819 painting by Caspar David Friedrich, and named the work 'The Unknowable'.[202] Conway sued for copyright infringement, and obtained an injunction on preliminary proceedings.[203] However, in the main proceedings, the Landgericht Berlin declared

188 s 30A CDPA.

189 Case C-201/13 *Deckmyn v Vandersteen* ECLI:EU:C:2014:2132, [2014] E.C.D.R. 21. Parody has two elements, 'first, to evoke an existing work while being noticeably different from it, and, secondly, to constitute an expression of humour or mockery.'

190 Case C-201/13 *Deckmyn v Vandersteen* ECLI:EU:C:2014:458, Opinion of AG Cruz Villalón, para 46.

191 See also Jonathan Griffiths, 'Fair Dealing after Deckmyn - The United Kingdom's Defence for Caricature, Parody or Pastiche' in Megan Richardson and Sam Ricketson (eds), *Research Handbook on Intellectual Property in Media and Entertainment* (Edward Elgar 2017).

192 *Deckmyn* (n 189) para 33.

193 Oxford English Dictionary Online (OUP March 2023) <https://www.oed.com/view/Entry/27973> accessed 25 November 2023.

194 Griffiths (n 191).

195 Andres Guadamuz, 'Artificial Inteligencie Parodies' (*TechnoLlama*, 2 July 2023) <https://www.technollama.co.uk/artificial-intelligence-parodies> accessed 25 November 2023.

196 ibid.

197 Not to be confused with how US fair use deals with the subject, see: Jiarui Liu, 'An Empirical Study of Transformative Use in Copyright Law' (2019) 22 Stanford Technology Law Review 163; and Gillotte (n 86).

198 See: Sotiris Petridis, 'Postmodern Cinema and Copyright Law: The Legal Difference Between Parody and Pastiche' (2015) 32 Quarterly Review of Film and Video 728.

199 Emily Hudson, 'The Pastiche Exception in Copyright Law: A Case of Mashed-Up Drafting?' [2017] Intellectual Property Quarterly 346, at 347.

200 Ruth Flaherty, 'Fair Dealing in a Fandemic: How Pastiche Can Be Used to Clarify the Position of User-Generated Content' (2022) 13 European Journal of Law and Technology 1 <https://www.ejlt.org/index.php/ejlt/article/view/877> accessed 25 November 2023.

201 Landgericht Berlin, 2 November 2021 – 15 O 551/19, <https://oj.is/2396832> accessed 25 November 2023.

202 Kate Brown, 'How Meme Culture and a Landmark Legal Case Against an Artist in Germany May Loosen Europe's Tight Copyright Regulations' (*artnet news*, 11 April 2022) <https://news.artnet.com/art-world/martin-eder-lawsuit-2091647> accessed 25 November 2023.

203 Kammergericht, 30 October 2019 – 24 U 66/19. More details about the decision here: Susan Bischoff, 'The dawn of pastiche: First decision on new German copyright exception' (*Kluwer Copyright Blog*, 7 June 2023) <https://copyrightblog.kluweriplaw.com/2023/06/07/the-dawn-of-pastiche-first-decision-on-new-german-copyright-exception/> accessed 25 November 2023.

that 'The Unknowable' is actually a work of pastiche, becoming one of the first applications of this exception in European courts. The court usefully defines pastiche as requiring 'an evaluative reference to an original […]. In contrast to illegal plagiarism, the older work must be used in such a way that it appears in a modified form. To do this, it is sufficient to add other elements to the work or to integrate the work into a new design,'[204]

The court usefully gave examples of possible pastiche uses, by stating that 'the pastiche in particular allows certain user-generated content (UGC) to be legally permitted […]. Quoting, imitating and learning cultural techniques are defining elements of intertextuality and contemporary cultural creation and communication in the 'Social Web'. In this context, practices such as remixes, memes, GIFs, mashups, fan art, fan fiction or sampling should be particularly considered.'[205]

At first glance, it would seem difficult to apply pastiche to an AI-generated output, but if we class such works in the same category as other UGC, and therefore in line with remixes and other similar practices, one could convincingly argue that generative outputs are a form of evaluative reference of an original work that is also a modified version of the original.

Taking both academic commentary and the German case as references, the transformative element of a work would be fair dealing if it is sufficiently different to the original, but similar enough to evoke it. This could potentially cover some of similar works, like the ones pictured above: you can tell that one depicts Superman because he's wearing his distinctive suit and has iconic black hair, but he is in an unfamiliar situation, and the face looks old and emaciated. As for the Pikachu picture, something like that could potentially work as pastiche, and even parody – the idea of a cute Pokémon at a bar is humorous, while still being recognisable.

Substantial copying is of course entirely determined on a case-by-case basis, so the discussion here is intended as a general analysis. Some generative uses could very well be found to infringe copyright, while others could either lack substantiality, or fall under the pastiche exception.

V. Conclusion

Every technological revolution comes with a considerable amount of legal strife. The early days of the World Wide Web, the days of filesharing and peer-to-peer, or the birth of user-generated content[206] all came with a few copyright infringement cases that sometimes defined the technology. Sometimes, those changes were unforeseen: Napster lost its case, but filesharing would continue unabated until legal downloads and music streaming became the norm. The early web infringement cases led to changes in the law so that intermediaries could operate without getting sued all the time for the actions of their users.

We are certainly living through a technological revolution in the shape of generative AI, and we do not know what shape it will take, but the lawsuits have already started flying in. We can expect a few years of strife, and perhaps at some point the law will become settled, either through the emergence of solid legal precedent, or by legislative action. We do not know the shape that those decisions will take. This article has tried to give an overview of the technology, but also of the law as it stands at the very early days of this revolution. Perhaps when reading this article back, it will be as outdated as those that were published in the early days of filesharing, useful only for historical reference.

One thing that could start happening now is for developers and rightsholders to begin looking at possible technological solutions to some of the challenges faced by the deployment of artificial intelligence. Some technological common ground could emerge, such as the widespread adoption of technical metadata standards that could flag a creator's wish not to have their works used in training. Private proactive initiatives such as Holly Herndon's opt-out website[207] are also steps in the right direction, empowering rightsholders to have more control. Similarly, technological solutions such as that developed by Vyas et al.[208] could be used to ensure that AI models do not regurgitate memorised inputs.

It is easy to use trite references to genies, bottles, cats, bags, and boxes when talking about artificial intelligence, but in this case, it may be apt to do so. The genie is indeed out of the bottle and the law will have to respond. Perhaps what is left is to negotiate the three wishes with the genie. Let us hope for a resolution that allows rightsholders to have equitable solutions, while also allowing the innovation of tools that could enhance our lives in positive ways.

204 LG Berlin 2 November 2021 at 40.

205 ibid at 39.

206 For example, *United States v LaMacchia* 871 F.Supp. 535; *MGM Studios v Grokster* 545 U.S. 913; and *A&M Records, Inc. v Napster, Inc.* 239 F.3d 1004 (9th. Cir., 2001).

207 <https://haveibeentrained.com/> accessed 25 November 2023.

208 Vyas (n 170).