Getty Images v Stability AI: Why Should UK Copyright Law Require Licences for Text and Data Mining Used to Train Commercial Generative AI Systems?

Zoya Yasmine
Somerville College, University of Oxford

In 2023, Getty Images commenced legal proceedings in the United Kingdom High Court against Stability AI. Getty Images claims that 7.3 million images from its database were unlawfully used to train Stability AI's generative Artificial Intelligence system. Drawing inspiration from Getty Images v Stability AI, this paper addresses the complexities surrounding copyright protection for text and data mining (TDM) in the UK. It argues that expanding Section 29(A) of the Copyright, Designs and Patents Act 1988 to exempt commercial AI developers from TDM licensing obligations would undermine the creative sector and hinder responsible innovation. This paper outlines the case's background and provides justifications for requiring TDM licences in the training of commercial generative AI systems. It argues that licensing requirements prevent the unjust appropriation of creators' work, foster valuable collaboration between creators and AI developers, and could even create new markets for existing works. The paper addresses practical challenges of TDM licensing, such as high costs, complexity, and the opacity of generative AI models. To address these issues, it proposes a set of reforms, including the adoption of standardised contracts for TDM, cross-licensing arrangements to facilitate fair data exchanges, and "nutrition labels" on AI-generated content to increase transparency and accountability. The paper concludes that these reforms, alongside the proposed court decision in *Getty Images*, could strengthen the UK's AI and art industries by promoting innovation within a fair legal framework that strikes an appropriate balance of rights between technology developers and creators.

Keywords: Copyright; AI; Licensing; Text And Data Mining; Generative AI

Introduction

In 2023, Getty Images commenced legal proceedings in the United Kingdom High Court against Stability AI (Getty Images, 2023). Getty Images claims that 7.3 million images were unlawfully scraped from its website by Stability AI to train its Generative Artificial Intelligent System (GAIS) without an appropriate licence (Getty Images, 2023). The Copyright, Designs and Patents Act 1988 (the Act) provides Getty Images with copyright protection over its visual asset database, so unless an exception applies, permission (through a licence) is required if other parties wish to use or copy these images. Section 29(A) of the Act provides an exception which permits copies of any copyright protected material for the purpose of Text and Data Mining (TDM) without a specific licence if this is for *non-commercial purposes.*

TDM is the automated technique used to extract and analyse vast amounts of online text or data to reveal relationships and patterns in data (Holland, 2021). TDM has become an increasingly valuable tool to train lucrative and beneficial GAIS on mass amounts of data scraped from the Internet. But as profitable technology companies are using this process to train their GAIS without a licence, the Intellectual Property (IP) rights attached to training data have been under scrutiny because it is unclear whether developers need a TDM licence to train their *commercial systems* on copyright-protected materials. There has been a flood of copyright infringement cases against AI companies who have chosen not to use TDM licences to train GAIS, but most of these are against American companies in the US Courts (Lutkevich, 2024). *Getty Images v Stability AI* is the first case of its kind in the UK.

In 2022, the UK Government's Intellectual Property Office (IPO) proposed to broaden the scope of Section 29(A) to provide commercial generative AI companies, like Stability AI, with unprecedented access to train its systems on copyright-protected materials without a TDM licence (the Proposal). The Proposal was designed to align with the Government's (2021) National AI Strategy to make the UK the most attractive landscape for AI development and investment. AI developers claimed that GAIS would not exist without wide exceptions to copyright law which permit the free use of TDM on copyright-protected materials (Milmo, 2024). However, a few months after, the IPO was forced to pause its Proposal due to backlash from the creative industry who argued that their works should not be used as free training data without compensation provided by a TDM licence (House of Commons, 2023; Orlowski, 2024). It is unclear whether the Government plans to re-introduce the Proposal, but the IPO is likely awaiting the outcome of *Getty* to set the UK's future approach.

In this paper, I use the facts of *Getty Images v Stability AI* as a platform to consider how the judge should resolve this case. This paper argues that the IPO's Proposal (2024) overlooked the innovative and collaborative value of licensing in relation to the AI copyright "input dilemma". I propose that in relation to the upcoming case, the Court should decide in favour of Getty Images. This judgment would affirm the current scope of Section 29(A) so only entities using copyright-protected materials for *non-commercial purposes* will be able to do so without a TDM licence. There is a perception that requiring licences will stifle AI development and frustrate the Government's pro-innovation approach to AI regulation (Milmo, 2024). Throughout this paper, I argue that licences can encourage AI innovation, but also allow the creative industry to flourish.

The original contributions of this paper can be seen as threefold. Firstly, there is limited academic literature that clearly outlines the UK's copyright landscape in relation to TDM and GAIS. Academic commentary has focused on jurisdictions where there are more cases being decided based on this dilemma and the scale of AI development is larger – for example, the US, EU, or Japan (Dermawan, 2023; Manteghi, 2023; Li, 2024). In addition, beyond offering a descriptive account of the law, this paper also focuses on the normative, more ambitious, question: how *ought* UK copyright law apply to the training of commercial GAIS using unlicensed materials? I ground my analysis in *Getty Images* as an opportunity to consider the real implications and practicalities of these cases.

The second contribution is based on the interdisciplinary analysis that I adopt to reason how *Getty Images* should be decided. To date, lawyers, AI developers, and creatives, have been responding to this question in isolation. Thus, this paper aims to unify discourses between these communities and recommends a solution which balances the interests of the law, advancement of technology, and preservation of creatives' rights. Finally, this paper is also committed to go beyond description, analysis, and critique, by providing policy recommendations about how TDM licences can be improved to satisfy the needs of our growing technology industry and safeguard artists from copyright infringement. This reform-oriented element is seen as necessary to ensure that the benefits of the "law in books" translates into the "law in action" (Hutchinson, 2015). Focussing on the "law in action", I include real case studies and examples to point to opportunities to improve our TDM licensing landscape.

In Section I, I outline *Getty Images* and the UK legal framework that applies to TDM. I will then briefly outline how the Court should decide in favour of Getty Images. Sections II and III will focus on the benefits and challenges of requiring AI developers to seek a licence to train their commercial GAIS on copyright-protected materials. Section II will explore three justifications for maintaining the scope of Section 29(A). For the first justification, I argue that a TDM licence is required so GAIS do not unfairly freeride off creator's content. The second justification argues that mandating TDM licensing will encourage creators and AI developers to unlock untapped value in materials and prevent obstacles that stifle innovation. For my final justification, I dispute claims that GAIS will erode the market for original works that serve as training data for GAIS – I suggest that TDM licensing could spur a new demand for existing works.

In Section III, I acknowledge that despite these justifications, issues with TDM remain, namely concerning the: (a) cost, (b) complexity, and (c) opaqueness of GAIS. Last year, the IPO announced that it was establishing a Code of Practice (COP) to improve the TDM licensing environment (IPO, 2023; Foerg, 2023). Just a few months after this announcement, the COP was abandoned as members of the committee could not agree on policies that balanced the rights of the creative industry and AI developers (Thomas and Criddle, 2024). In the final section of this paper, I respond to the challenges set out in Section III and provide some measures that could mitigate the shortfalls of TDM licences that should be implemented by the IPO. As this paper is mainly dedicated to the UK's *legal* response, the suggestions for the IPO are brief and provided as a platform for further research to supplement the proposed Court decision.

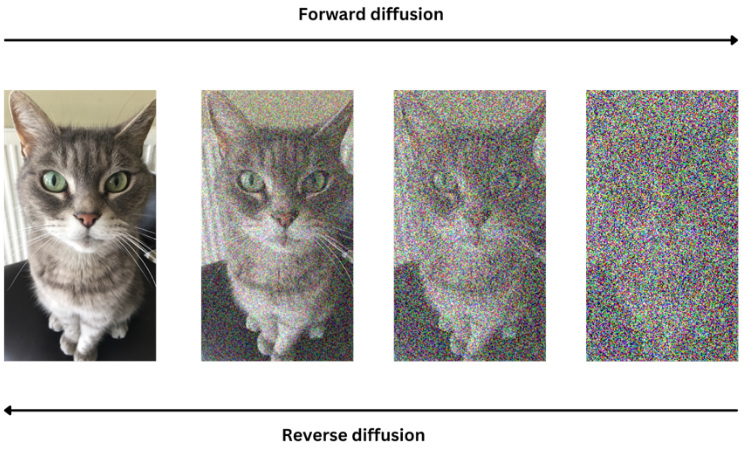
1. How Should the UK High Court Decide *Getty*?

In this Section, I outline the technical elements of *Getty* and map out the current UK copyright law in relation to TDM under Section 29(A) of the Act. I then argue that Stability AI infringed copyright when the company used Getty Images' protected materials to train its GAIS (called Stable Diffusion) without a TDM license.

1.1. Getty Images v Stability AI

To assess whether Stability AI violated Getty Images' copyright, it is important to understand how Stable Diffusion, an AI tool that turns text into images, is trained. This process involves utilising images from various online databases, including Getty Images, but these images are not stored directly. Instead, the AI developers utilise a specific training method, like a "diffusion model", to enable the model to learn patterns from the images.

The "diffusion model" training process works by adding random visual "noise" to each of the image present in the training dataset until the image is not recognisable – this process is understood as "forward diffusion" (Guadamuz, 2024). Once the images are "noised", the AI is trained to recognise and gradually remove that noise to reconstruct the original image in a process known as "reverse diffusion" (Guadamuz, 2024). Figure 1 illustrates the "diffusion model" training process using an image of a cat as an example. Through repeated training on thousands of images, the AI model learns to identify patterns, like what common objects and colours look like. As a result, the GAIS can start to generate new images based on these learned patterns.



Overlay image credit: Adrien Limousin / Better Images of AI / Non-image / CC-BY 4.0

Figure 1: *Illustrates the noising process during the diffusion model training process for an image of a cat*

Importantly, the images generated by the AI model in its output will not be exact copies of any original images used in the training process. Instead, the outputs are statistical approximations learned during the training process which inform the model's overall understanding of how objects are represented (Guadamuz, 2024). Getty Images' extensive library of over 12 million images served as a rich resource for training data for GAIS, contributing to Stable Diffusion's enhanced ability to generate vast, realistic outputs.

Copyright law becomes relevant in this training process when we focus on what this framework aims to protect. Copyright law determines that the protected element of works subsides in the creative expression – like the lighting, exposure, filter, or positioning of an image (*Temple Island Collections*, 2012). These are the parts of images that copyright protects because they require a creator's own thoughts and originality. However, what is significant about the GAIS training process for copyright law is that Stability AI does not use TDM to copy Getty Images' database for the protected elements of its materials (Lemley and Casey, 2020).

To train GAIS, it is often the *factual* elements of the work extracted through TDM which are the most valuable as opposed to the creative aspects (Lemley and Casey, 2020). The diffusion model training process relies on broad visual features of images, rather than specific artistic choices. For example, when training Stable Diffusion, TDM was not used to extract data about lighting techniques which were employed to make an image of a cat particularly appealing. Instead, the accessibility to a large collection of images which detailed the features that resemble a cat (fur, whiskers, big eyes, paws) were what Getty Images' database provided. The challenge for Stability AI is that it is unable to capture these unprotectable parts of the images that are essential for training Stable Diffusion, without making a copy of the protectable parts (Lemley and Casey, 2020).

1.2. The current UK copyright law framework in relation to TDM

In Section 29(A) of the Act, the UK currently permits TDM of copyrighted works for *non-commercial* purposes provided that the entity has lawful access to the work. Lawful access means that individuals do not require separate permission for TDM, they just require access to the works through a general licence or subscription (IPO, 2014). Section 29(A) is also mandatory, so even if contract terms to access materials might preclude TDM, these are unenforceable (IPO, 2014). Given that academics and researchers often have broad institutional access to materials, the UK Government has exercised a very facilitative approach to TDM for training non-commercial GAIS to drive scientific advancements (Flynn and Vyas, 2023).

The question of whether the training of Stability AI classifies as a *non-commercial* purpose is likely to be an unproblematic for the Court because: (i) it has already been decided that it is a commercial entity in the US case (*Getty*, 2023), (ii) Stable Diffusion was monetised (*ibid.*), and (iii) the Government intended for Section 29(A) to be used by universities and charities (IPO, 2014). I acknowledge that UK data laundering practices (where commercial technology companies outsource data collection and model training to academics) present a loophole in this framework that must be addressed (Baio, 2022), but consideration of this is beyond the scope of this paper given that this did not occur in *Getty.* Therefore, Stable Diffusion will likely fall outside the non-commercial exception in Section 29(A). It will be for the Court in *Getty* to decide whether to extend Section 29(A) to *commercial* use (as in the Proposal) to free Stability AI from copyright infringement.

1.3. How should the UK's High Court decide Getty?

I argue that the Court should decide in favour of Getty Images and refrain from expanding Section 29(A) to commercial GAIS. Therefore, Stability AI infringed copyright when it did not acquire a TDM licence to train its system on Getty Images' protected materials.

It is important to note here that since Getty Images' legal action in 2023, Stability AI has later filed a defence against its copyright infringement (Cooke, 2024). Stability AI is arguing that it cannot be held liable for copyright infringement in the UK because the training of its GAIS took place on servers in the US (*ibid*.). Stability AI originally tried to have the case struck out based on this jurisdictional fact. However, the judge overseeing the litigation decided that the case should go to trial so more evidence could be gathered about this matter (Davies and Dennis, 2024).

Thus, there remains a strong possibility that Stability AI's defence will not be upheld in court and the judge will have to determine how the scope of Section 29(A) applies to the case (*ibid.*). It is also possible that the judge will address this question of law in the case regardless of the jurisdiction in which the training took place. In a recent AI and patent case (*Emotional Perception*, 2023), the judge went beyond resolving the matters between the parties to answer wider questions of law relating to the patenting of artificial neural networks (*ibid.*). It is assumed that this is because of the long backlog of cases and the rapidly evolving development of AI which requires faster responses and legal certainty to protect creators and AI developers. Therefore, a decisive ruling in *Getty Images v Stability AI* is welcomed to provide much-needed legal guidance to the industry and to align UK copyright law with the rapid development of commercial GAIS.

2. Justifications For Requiring TDM Licences To Train Commercial GAIS

In this Section, I provide three justifications for my proposal aimed at balancing the IP rights of original creators with the goal of fostering AI innovation according to the Government's Strategy (2021). Firstly, I suggest that requiring TDM licences means that generative AI developers cannot freeride off creator's works. The freeriding argument claims that creatives will lack sufficient incentives to develop new works if their materials are leveraged by others without fair compensation (Lemley, 2005). Despite claims that the freeriding argument does not apply to GAIS, its relevance persists when considering how the value of existing works can be reimagined when used as training data. Secondly, TDM licences can enable a more collaborative innovation process, supporting AI developers in creating advanced GAIS more efficiently. Finally, contrary to the perception that GAIS will diminish market value for original works, I propose that mandating TDM licences might occasionally reinvigorate demand for creators' original works.

2.1. Freeriding and reimagined value

Protectionist IP theorists argue that copyright law should uphold a robust exclusionary right to prevent unauthorised use of protected works (Lemley, 2005). Getty Images (and its photographers) invest substantial resources in curating a high-quality image repository, with over $200 million invested between 2017 and 2020 alone (Getty Images, 2023). Photographers depend on the royalties received from Getty Images to sustain their livelihoods and continue producing content (Getty Images, 2023). Copyright protection thus enables Getty Images to maintain profitability by determining its competitors from using its images without bearing the associated costs of production. Without fair remuneration, Getty Images and its contributors would not have the resources, incentives, or time to invest in its database. To date, Stability AI has raised more than $100 million in financing (Getty Images, 2023). But without scraping images from Getty Images' database, Stability AI might not have had access to the extensive data needed to train its model effectively. The success of Stable Diffusion rests on the time and investment of Getty Images (and its photographers) into its database. Stability AI's reluctance to seek a licence amounts to freeriding on Getty Images' materials. Therefore, mandatory TDM licences will ensure that commercial GAIS cannot benefit from protected works without compensation to the creator to recognise how these materials are the foundation of GAIS.

The freeriding argument has been criticised in relation to the training of GAIS (Lemley and Casey, 2020). This is because, as explored in Section I, TDM does not extract the protected elements of copyright materials. According to this argument, Stable Diffusion does not freeride on Getty Images' photograph of a cat. The factual elements that compose a cat are not connected to a photographer's time and investment into the image – this is only directed at the expression of the cat (captured in the angle of a shot, exposure, or colour manipulation) which Stable Diffusion did not capitalise on (Lemley and Casey, 2020). However, the potential for TDM to "re-imagine" the value of such materials suggests that the freeriding argument may still apply.

An example of re-imaged value is demonstrated by Gmail's predictive email response algorithm which was trained on romance novels (Smith, 2016). Google leveraged the fact that these romance novels would provide convenient training data for its algorithm to learn varied language, phrasing, and grammar structures. The algorithm was not used to replicate specific story elements like the characters, settings, or descriptive tone. Instead, its sole use was for the purpose of understanding the English language (Smith, 2016). Nevertheless, these romance novels were still valuable (albeit in a reimagined way) to the success and effectiveness of Gmail's tool.

Similarly, Stability AI's use of Getty Images' database illustrates how re-imaged uses can result from TDM practices. It would have been significantly more difficult for Stability AI to train its GAIS without the convenience, existence, and volume of data extracted from Getty Images' vast database. Thus, even though this is not connected to the traditionally protected elements of images under copyright law per se, the underlying freeriding motive still stands. The use of TDM to train GAIS still freerides on the creator's materials by extracting valuable data from existing materials which would not exist without creators' significant time, resources, and efforts.

I do not suggest that the boundaries of copyright law should be extended to protect all materials that serve as the basis of profitable innovation. Copyright law maintains appropriate exceptions to protection for scientific formulas or symbols to ensure the necessary access to the basis of our scientific and creative developments. However, I do argue that copyright law should reassess what was traditionally deemed unprotectable in light of GAIS to ensure that the law still supports the appropriate balance of rights. In this context, TDM licences could ensure that AI companies appropriately compensate creators for their works which provide the foundations of profitable GAIS.

2.2. Collaboration to unlock untapped value

A central feature of the IP system is the licensing framework, which enables lawful access to copyright protected materials to progress innovation. Therefore, the fact that TDM can extract untapped value in existing materials to develop new and innovative AI systems is exactly what copyright law supports (Leval, 1990). Examples of untapped value include user interactions on social media being used to train virtual assistants (Meta, 2023) and international legislation texts used to train deep learning translation tools (DeepL, 2023). These uses illustrate how existing, protected content can contribute significantly to the development of further innovation, like GAIS. Copyright law stands to incentivise individuals to develop upon existing protected works using licences to unlock further creations which are socially beneficial.

Thus, the second reason that copyright law should encourage TDM is because it saves AI developers time and resources from training systems when resourceful data already exists. AI innovation efforts can then be directed at developing cutting-edge GAIS, as opposed to data creation and training. A legal framework that offers clarity on IP rights related to training data could encourage creators and AI developers to explore usually beneficial uses of existing content (Brook and Murray-Rust, 2014). TDM licences would allow creators to profit from such uses, while fostering a collaborative environment that strengthens the development of GAIS.

Stability AI had already started to re-imagine the use of existing materials by leveraging Getty Images' database which was originally designed for use by media and corporate companies. However, since Stability AI did not obtain a TDM licence, the materials scrapped from Getty Images' website were low-quality and distorted by watermarks (Getty Images, 2023). A formal licensing agreement would have enabled access to high-quality data, and might have also encouraged collaborative enhancements, including machine-readable metadata which would have streamlined and enhanced the training process. Getty Images have already worked with AI companies, so licensing negotiations could have also offered opportunities for Getty Images to further improve Stable Diffusion's development process with its valuable domain knowledge and experience (Getty Images, 2023). Thus, TDM licences facilitate collaboration between AI developers and creators which is necessary to better optimise training data to efficiently develop better GAIS.

Without adequate compensation measures provided by TDM licences, creators are stifling the innovation process (Shan *et al.*, 2023). Using data tags on their materials (like robots.txt which contain do-not-scrape directives to block web crawlers), creators are blocking and distorting the TDM processes to retain control over their works. Data tags, like Nightshade, can even "poison" the TDM process by sending back the incorrect images to distort the accuracy of GAIS's training process (Shan *et al.*, 2023). The use of data tags has been an act of resistance from creatives against AI companies freeriding on their materials. But this is not because creators are reluctant to have their works being used as training data *per se*; creators just want to control the use of their works and ensure that they are adequately compensated (Dean, 2023).

Data tags and other resistance efforts create a divergence between AI companies and creators, preventing any possibility of their works being used for remuneration and corrupting the training process for GAIS. Furthermore, as materials are being increasingly withheld from AI companies, this will ultimately lead to the self-demise of GAIS. New data is needed for GAIS to meet the evolving demands of consumers. Therefore, TDM licences offer a way to resolve the tensions between these two communities and support a more productive innovation process for GAIS whilst adequately compensating artists.

2.3. Re-invigorating value in original works

It is argued that the use of creator's materials as training data for GAIS will devalue the market for the original work (Sobel, 2017; Lucchi, 2023). An alternative hypothesis is that TDM could also hold the potential to occasionally *improve* the market for original works despite their inclusion in datasets for training GAIS. To explore this argument in a different context, Snapchat has made licensing agreements with minority artists to prompt users to use their music in videos. Snapchat has benefited from a cheaper method to obtain music on its platform and smaller musicians have benefited from increased exposure of their works on the popular app (Malik, 2022). While the original intention for these artists was not to produce works for this purpose, it provides an alternative avenue to attract audiences and generate additional market access.

In a similar way, using existing materials to train GAIS could actually prompt renewed appreciation for these works. Benn argues that AI art might increase the public's appreciation for human creativity, as human-centred works can carry emotional or aesthetic value that digital creations may not fully replicate (Aesthetics for Birds, 2022). Therefore, if a photographer or artist exclusively licences their unique database of images which are distinctive with respect to the style or skills needed to replicate the images, the licensing AI company will benefit from a significant competitive advantage.

Greg Rutowski is a Polish digital artist who uses classical painting styles to create fantasy landscapes which are used in illustrations for games like Dungeons & Dragons. His images have become more popular since his images were used as training datasets for text-to-image AI generators. Rutkowski was optimistic that this could be a good way to reach new audiences who appreciate and value his fantastical and ethereal artistic style. However, the problem is that the GAIS did not disclose or acknowledge the artists or sources for which the training materials were derived from so it was impossible for users to find Rutowski's artworks. Therefore, it is acknowledged that the strength of the "reinvigoration" argument relies on GAIS being transparent about their training materials, but also only where datasets hold certain unique value. But it is maintained that in these instances, TDM licences could drive revenue and appreciation towards the original materials.

3. Problems with TDM Licensing and Mitigating Measures for the IPO

In this Section, I outline three drawbacks with TDM licensing: cost, complexity, and opacity. While these problems raise valid concerns, I detail mitigating measures which could be implemented by the IPO to improve TDM licensing through industry changes.

3.1. Cost: cross-licensing arrangements

The main problem with TDM licences is that they are very costly for AI developers. GAIS require vast amounts of data to produce good quality outputs – just training the first two versions of Stable Diffusion required around 12 million images (Getty Images, 2023). Collating smaller datasets from individual owners is usually a time-consuming and expensive task (Lemley and Casey, 2020). Alternatively, the possibility of acquiring large datasets from bigger companies is unlikely as these have significant commercial value so are priced highly or not licensed at all. The BBC has admitted that it relies on its own proprietary data as licensing third-party materials for their AI tools is too expensive (BBC, 2022). The BBC is in a fortunate position to at least have its own data, but for smaller companies the cost of TDM licensing creates barriers to enter the AI market. The cost of TDM licences creates monopolies in

AI development as only a few companies can afford to licence third-party datasets or have access to their own data to train GAIS (Lucchi, 2023).

While TDM licensing might be expensive, adapting existing cross-licensing mechanisms to copyright-protected data could be a useful mechanism to help smaller companies develop and train their own GAIS (Fershtman and Kamien, 1992). A cross-licensing agreement occurs where parties exchange licences (instead of money) for use of each other's IP. In this context, I suggest that companies with access to large (often homogenous) datasets could exchange their materials with smaller companies who may have more diverse datasets. Gaining richer data is important for companies to avoid their models' "overfitting" (creating outputs which replicate the training data) which could result in costly copyright claims in the output materials of GAIS (Carlini *et al.*, 2023). AI developers are also under increased pressure to limit the bias outputs of their GAIS – especially as new tools are being released to scrutinise unrepresentative models (Heikkilä, 2023).

An example of a cross-licensing opportunity could involve "Better Images of AI" licensing data about the accurate representation of AI in exchange for larger datasets from the BBC, giving each other the resources to generate valuable and representative GAIS. It is possible for terms in the cross-licensing agreement to stipulate that each party does not use the data for the same purpose, so they do not develop identical GAIS or saturate the market. I suggest that the IPO should raise awareness of TDM cross-licensing arrangements to reduce the monetary barriers required to enter the generative AI market and facilitate the creation of more diverse and cutting-edge AI tools.

3.2. Complexity: standardised licences

TDM licensing is also a time-consuming process if complex contracts are drafted which require legal assistance if parties want to have an informed understanding of the scope of data use for TDM (BBC, 2022; Vollmer, 2016). Big corporations can often leverage their powerful position to draft licences in an overly complex way to attain broad rights over creators' materials (Stevens, 2023; Sobel, 2017). To mitigate this problem, I suggest that the IPO creates standardised TDM (and even cross-licensing) contracts to be used between AI companies and creators. Standardised TDM licences will empower creators to licence their materials without the need to navigate the complex legal landscape to control the use of their data. Comprehensible licensing contracts will also streamline the innovation process for AI developers who can train GAIS faster without the need to spend time drafting contracts and negotiating TDM terms (Maffioli, 2023).

A global movement towards open-source standardised contracts for routine arrangements has already begun with Non-Disclosure Agreements (oneNDA, *n.d.*). While it is outside the scope of this paper to detail what standardised TDM contracts should include, Maffioli (oneNDA, *n.d.*) has proposed a standardised template that could be a good baseline for the IPO to develop. This proposed TDM contract includes terms relating to usage and access rights, risk allocations and liabilities, transparency provisions and compensation (oneNDA, *n.d.*). Therefore, the complexity of TDM licensing could be mitigated if the IPO designs standardised TDM contracts for use between AI companies and creators.

3.3. Opacity: nutrition labels

Due to the vast amounts of data that GAIS are trained on, and the number of parameters within models, GAIS often produce content that does not resemble its training data which makes it difficult for creators to know if their materials have been unlawfully used as training data (Guadamuz, 2024). It is also in the best interests of the company to ensure that there is no resemblance with creator's works to avoid copyright claims targeted at the output imagery (Guadamuz, 2024). The images used to train Stable Diffusion were watermarked, so Getty Images could identify its images in the output imagery. However, images will not always be watermarked, so creators will be unaware of the use of their works as training data for GAIS. This creates a loophole for AI developers who can avoid obtaining TDM licences (even if legally required) because the opacity of GAIS provides a shield against accountability for the infringement of protected materials. Thus, TDM licensing is only effective if AI developers are forthcoming about their use of protected materials, or creators are made aware of the use of their works as training data for GAIS.

To increase transparency, I propose that the IPO implements a requirement for AI developers to embed "nutrition labels" on content created by GAIS (Lucchi, 2023; Maffioli, 2023). Nutrition labels are already being used by leading AI companies to disclose information about what data was used to create AI-generated images (Swant, 2023). By integrating nutrition labels onto output imagery, transparency is instilled in GAIS development, empowering creators to better recognise potential copyright infringements by GAIS and encouraging AI developers to scrutinise the origins of their training materials (Maffioli, 2023).

I do acknowledge that there are limits to transparency, as AI companies should not be expected to publicly disclose their training datasets or open-source their models – such would undermine a company's competitive advantage. However, in light of the opaqueness of models, creators should be afforded with greater awareness of whether their works are being unlawfully used as training data. Additionally, the requirement for nutrition labels aligns with the argument made in Section II which recommends that transparency could increase demand for creators' original works. From a commercial standpoint, GAIS with accredited sources are also perceived as more reliable and responsible by users (Swant, 2023). Thus, the IPO should mandate developers to embed nutritional labels on AI content to strike an appropriate balance between AI developers and creators, while promoting the advancement of improved GAIS.

Conclusion

This paper has argued that without a TDM licence, training commercial GAIS on copyrighted materials should be considered as infringement. In *Getty,* the Court should refrain from expanding the scope of Section 29(A) so only entities using copyright-protected materials for *non-commercial* purposes can do so without a TDM licence. Three reasons have been presented to highlight how the Proposal overlooked the benefits of TDM licensing for the AI and creative communities.

A majority of literature attempting to resolve dilemmas in the intersection of copyright law and AI do not focus on the UK jurisdiction and are not interdisciplinary in their analysis. In this paper, I attempted to address this research gap by focusing on the upcoming *Getty* decision as well as exploring reasons for deciding the case which balances the interests of the law, AI developers, and creatives. While I have focused on the UK, the justifications, challenges and recommendations outlined in Sections II and III can be adapted to other jurisdictions – especially where the courts have already ruled that TDM licences are necessary. The paper's more novel and optimistic perception of the value of TDM licensing will be helpful to bridge innovation efforts between the AI industry and creators. I hope that the hypothetical examples and real examples included in this paper shed light on how these communities can work together in a responsible and mutually beneficial way. Within the art industry, original creators, AI start-ups, minority artists, and large AI companies can all bring something to the innovation ecosystem if they want to. I challenge these actors to take collaboration opportunities more seriously and think about how they can use the law to facilitate this process to ensure their respective needs, commitments, and rights are upheld.

The solution to the copyright problems in relation to training commercial GAIS is complex. In this paper, I have been a strong advocate for the use of TDM licences as their innovation effects have often been overlooked. The proposals outlined in the final section of this paper serve as a purpose to show that while TDM licences can resolve some problems relating to freeriding and creator resistance, they are not perfect and require shaping to meet the demands of the working industry. While I have pointed to some of the shortfalls and mitigating measures, including standardising licensing, prompting cross-licensing opportunities, and utilising nutrition labels, further research is required. I suggest that further research adopts a more empirical methodology to investigate the real challenges relating to "licensing in action" faced by AI developers and creators. In this paper, I used the Government's consultation on IP and AI which yielded responses from various actors with different interests, like the BBC, IBM, the Music Publishers Association, The Law Society, Siemens, and the Wellcome Trust to name a few (Intellectual Property Office, 2022). However, given that these all groups submitted to the consultation, the responses might not represent wider views in the ecosystem from underrepresented artists and smaller AI developers who might also be facing issues that have not been reported or raised. I hope that the recommendations provided in this paper can set out the first steps for other researchers to advocate for changes to our licensing and innovation frameworks to protect creators and improve clarity over the scope of rights in the face of GAIS.

Finally, in relation to the wider intersection between copyright and generative AI, this paper has exclusively focussed on the "input question". But questions remain to be answered in relation to whether the outputs of GAIS can infringe on creators' copyright. The TDM licensing approach suggested in this paper is one way to facilitate a better dynamic between the AI and creative industry which could limit the legal action necessary to monitor the output imagery by resolving issues in initial licensing negotiations. For instance, TDM licences could allow AI developers and creators to negotiate in advance to compensate artists if the outputs of GAIS are to the likeness or similarity of the artist's original work. Future research could understand how TDM licences, if at all, could benefit legal questions focussed on infringement of the output imagery. This would provide a more rounded and comprehensive understanding of the innovative value of TDM licences in relation to GAIS

References

Aesthetics for Birds, (2022). Eights Scholars on Art and Artificial Intelligence. *Aesthetics for Birds,* 2 November. Available at: https://aestheticsforbirds.com/2023/11/02/eight scholars-on-art-and-artificial-intelligence/. (Accessed: 30th November 2023).

Andersen v. Stability AI Ltd (2023). U.S. District Court for the Northern District of California, No. 3:23-cv-00201.

BBC (2022). *BBC Response to the UKIPO Consultation on AI and IP: Copyright and Patents.* Available at: https://www.gov.uk/government/consultations/artificial-intelligence and-ip-copyright-and-patents (Accessed: 30th November 2023).

Baio, A. (2022). AI Data Laundering: How Academic and Non-profit Researchers Shield Tech Companies from Accountability. 30 September. Available at: https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield tech-companies-from-accountability/. (Accessed: 14 January 2024).

Brook, M., Murray-Rust, P. & Oppenheim, C. (2014). The Social, Political and Legal Aspects of Text and Data Mining (TDM). *D-Lib Magazine* (November). Available at: https://openaccess.city.ac.uk/id/eprint/4784/1/D-Lib%20Magazine.pdf.

Carlini, N *et al.* (2023). Extracting Training Data from Diffusion Models. Available at: https://arxiv.org/abs/2301.13188.

Cooke, C. (2024). Stability files defence in important test case on AI and UK copyright law. *Complete Music Update.* Available at: https://completemusicupdate.com/stability-files defence-in-important-test-case-on-ai-and-uk-copyright-law/. (Accessed 4 March 2024)

Copyright, Designs and Patents Act 1988, Section 29(A). Available at: https://www.legislation.gov.uk/ukpga/1988/48/section/29A (Accessed: 30th November 2023).

Davies, C. Dennis, G. (2024). Getty Images v Stability AI: the implications for UK copyright law and licensing. Pinsent Masons, 29 April. Available at: https://www.pinsentmasons.com/out-law/analysis/getty-images-v-stability-ai-implications copyright-law-licensing. (Accessed 19 June 2024)

DeepL (2023). Why AI translation is a must-have for legal firms with global colleagues and clients. DeepL, 11 April. Available at*: https://www.deepl.com/en/blog/why-ai-translation-is a-must-have-for-legal-firms-with-global-colleagues-and-clients.* (Accessed: 11 April 2023).

Dermawan, A. (2023). Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese "nonenjoyment" purposes. *The Journal of World Intellectual Property* 27(1).

Emotional Perception AI Ltd v Comptroller-General of Patents, Designs and Trade Marks [2023] EWHC 2948 (Ch).

Fershtman, C Morton, I. (1992). Cross licensing of complementary technologies. International Journal of Industrial Organisation, 10(3).

Flynn, S. Vyas, L. (2023). Examples of Text and Data Mining Research Using Copyrighted Materials. *Kluwer Copyright Blog*, 6 March*.* Available at: https://copyrightblog.kluweriplaw.com/2023/03/06/examples-of-text-and-data-mining research-using-copyrighted-materials/. (Accessed: 20th January 2024).

Foerg, M. (2023). The UK government steps towards a code of practice on copyright and AI. *Kluewer Copyright Blog,* 27 September. Available at: https://copyrightblog.kluweriplaw.com/2023/09/27/the-uk-governments-steps-towards-a code-of-practice-on-copyright-and-ai/. (Accessed 19 June 2024).

Getty Images v Stability AI. (2023). United State District Court of Delaware. No. 1:23-cv 00135-UNA.

Getty Images. (2023). *Getty Images Statement.* Available at: https://newsroom.gettyimages.com/en/getty-images/getty-images-statement (Accessed: 30th November 2023).

Guadamuz, A (2024). A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs. *GRUR International,* 140(2).

Heikkilä, M. (2023). These new tools let you see for yourself how biased AI image models are. *Technology Review,* 22 March. Available at: https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for yourself-how-biased-ai-image-models-are/. (Accessed: 30th November 2023).

Holland, C. (2021). Copyright and Text & Data mining – what do I need to know?. *Open@UCL Blog,* 6 July. Available at: https://blogs.ucl.ac.uk/open access/2021/07/06/copyright-text-data-mining/ (Accessed: 30th November 2023).

House of Commons (2023). *Artificial Intelligence: Intellectual Property Rights.* Available at: https://hansard.parliament.uk/commons/2023-02-01/debates/7CD1D4F9-7805-4CF0-9698-E28ECEFB7177/ArtificialIntelligenceIntellectualPropertyRights (Accessed: 30th November 2023).

Hutchinson, T. (2015). The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law. *Erasmus Law Review* 3.

Intellectual Property Office (2014). *Exceptions to Copyright.* Available: https://www.gov.uk/guidance/exceptions-to-copyright. (Accessed: 30th November 2023).

Intellectual Property Office (2022). *Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation.* Available: https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents government-response-to-consultation. (Accessed: 30th November 2023).

Intellectual Property Office. (2023). *The government's code of practice on copyright and AI.* Available at: https://www.gov.uk/guidance/the-governments-code-of-practice-on-copyright and-ai (Accessed: 30th November 2023).

Leffer, L. (2023). Your Personal Information Is Probably Being Used to Train Generative AI Models. *Scientific American*, 19 October. Available at: https://www.scientificamerican.com/article/your-personal-information-is-probably-being used-to-train-generative-ai-models/. (Accessed 30th November 2023).

Lemley, M. (2005). Property, Intellectual Property, and Free Riding. *Texas Law Review* 83(291).

Lemley, M. Casey, B. (2020). Fair Learning. *Texas Law Review* 99(4).

Level, P. (1990). Towards a Fair Use Standard. *Harvard Law Review* 103(5).

Li, J. (2024). Managing Copyright Infringement Risks in Generative Artificial Intelligence Data Mining' *4th International Conference on Management Science and Industrial Economy Development* 39.

Lutkevich, B (2024). AI lawsuits explained: Who's getting sued?. *Tech Target,* 2 January. Available at: https://www.techtarget.com/whatis/feature/AI-lawsuits-explained-Whos-getting sued (Accessed: 13th January 2024).

Maffioli, D. (2023). Copyright in Generative AI training: Balancing Fair Use through Standardization and Transparency. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4579322.

Malik, A. (2022). Snap's new creator fund will award independent musicians up to $100,000 per month. *Tech Crunch,* 28 July. Available at: https://techcrunch.com/2022/07/28/snaps new-fund-award-independent-musicians-100000permonth/guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAKgtyuQ5DB9lnJXeUMSLXttk4EGx_DRWjEGmlTNLk33nb7i8YYBi4lqX_Qg9_kE_naZiTObBUj4jfJwHnrEDC7eYADj3w96HAIETgOZWjtlOs4Y2jsPnCciVoDD0reGgmgBsyroNS3trQvYsRNsn34FXLzOsE5-q2I9TvIEIYa. (Accessed: 14 January 2024).

Manteghi, M. (2024). Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?. *Law, Innovation, and technology* 1.

Meta, (2023). Privacy Matters: Meta's Generative AI Features. *Meta Newsroom,* 27 September. Available at: https://about.fb.com/news/2023/09/privacy-matters-metas generative-ai-features/. (Accessed: 30th November 2023).

Milmo, D. (2024) ''Impossible to create AI tools like ChatGPT without copyrighted material, OpenAI says', *The Guardian,* 8 January. Available at: https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material openai (Accessed: 14 January 2024).

New York Times v Open AI (2024). United States District of Southern New York. No. 1:23-cv 11195.

OneNDa. (n.d.). *oneNDA started off as a bit of a whim which went global – overnight.* Available at: https://www.onenda.org/about (Accessed 30th November 2023).

Orlowski, A. (2024). How 'big tech' barons are plotting to steal Britain's creativity. *The Telegraph,* 28 October. Available at: https://www.telegraph.co.uk/business/2024/10/28/how-big-tech-barons-plotting-steal-british-creativity/ (Accessed 30 October 2024).

Rouse, M. (2024). Generative AI. *Techopedia,*15 January. Available at: https://www.techopedia.com/definition/34633/generative ai#:~:txt=Generative%20AI%20(genAI)%20is%20a,with%20the%20sam%20statistical%2 0properties (Accessed: 20th January 2024).

Shan, S *et al.* (2023). Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. *Cryptography and Security.* Available at: https://arxiv.org/abs/2310.13828.

Smith, L. (2016). Google's AI engine is reading 2,865 romance novels to be more conversational. *The Verge,* 5 May. Available *at:* https://www.theverge.com/2016/5/5/11599068/google-ai-engine-bot-romance-novels. (Accessed: 30th November 2023).

Snow, J. (2018). "We're in a diversity crisis": cofounder of Black in AI on what's poisoning algorithms in our lives. *Technology Review,* 14 February. Available at: https://www.technologyreview.com/2018/02/14/145462/were-in-a-diversity-crisis-black-in ais-founder-on-whats-poisoning-the-algorithms-in-our/ . (Accessed: 30th November 2023).

Sobel, B. (2017). Artificial Intelligence's Fair Use Crisis. *Columbia Journal of Law and the Arts* 41(45).

Stevens, M. (2023). If Big Brands Copied Their Work, What Are Artists to Do? *New York Times,* 2023. Available at: https://www.nytimes.com/2023/03/01/arts/design/digital-art copyright-marvel-panini-wizards.html. (Accessed: 30th November 2023).

Swant, M. (2023). Adobe debuts new icon as a 'nutrition label' for generative AI content'. *DigiDay,* 11 October. Available at: https://digiday.com/media-buying/adobe-debuts-new icon-as-a-nutrition-label-for-generative-ai-content/. (Accessed: 14th January 2024).

Temple Island Collections Ltd v New English Teas Ltd. (2012). EWPCC 1.

Thomas, D. Criddle, C. (2024). UK shelves proposed AI copyright code in blow to creative industries. *Financial Times,* 5 February. Available at: https://www.ft.com/content/a10866ec 130d-40a3-b62a-978f1202129e. (Accessed: 19 June 2024).

Vollmer, T. (2015). More licences are not the solution for text and data mining. *Communia,* 11 June. Available at: https://communia-association.org/2015/06/11/more-licenses-are-not the-solution-for-text-and-data-mining/. (Accessed: 30th November 2023).