Published on 29 November 2023

GENERATIVE AI AND COPYRIGHT

Part 1: Copyright Infringement

[2023] SAL Prac 24

The ability of artificial intelligence ("AI") to autonomously, or in response to specific prompts by human users, generate text, music and images, is developing at a breathtaking pace. Generative AI applications such as ChatGPT, DALL·E and Stable Diffusion are presenting new scenarios that can be confounding to copyright lawyers as well as judges and policymakers. This two-part article discusses whether existing copyright doctrines can adequately address these issues and to what extent a recent overhaul of the Copyright Act in Singapore in 2021 is equipped to deal with these challenges. Part 1 deals with copyright infringement in both the input of works to train generative AI applications and the AI output in response to human prompts and commands. Part 2 discusses how the computational data analysis exception and open-ended fair use provision in the Copyright Act 2021 are likely to apply to these scenarios.

David **TAN**
Professor, Faculty of Law, National University of Singapore;
Co-Director, Centre for Technology, Robotics, Artificial Intelligence & the Law;
Head (Intellectual Property), EW Barker Centre for Law & Business.

I. Introduction

1 Globally, and in Singapore, there is certainly significant public interest in what ChatGPT can deliver, whether in assisting students with writing school assignments or in generating scam e-mails. ChatGPT – where "Chat" in the name refers to it being a chatbot, and "GPT" stands for generative pre-trained transformer which is a type of large language model ("LLM") – is the artificial intelligence ("AI") system designed by OpenAI which builds generative models using deep learning technology

that leverages large amounts of data to train an AI system to perform a task.[1] It has been reported to be the fastest-growing consumer application in history, far surpassing the successes of TikTok, Facebook and Instagram.[2] OpenAI also operates DALL·E which is an AI system that can create realistic images and art from a description in natural language. These sophisticated AI technologies which train on vast quantities of authorial works to generate new content in response to text prompts are often described as "generative AI",[3] and the manner in which these copyright-protected works are employed in training the AI has attracted a number of high-profile lawsuits since the start of 2023.[4] For the purposes of this article, the panoply of generative AI applications will be known as "GAIAs".

2 The temporary ban on ChatGPT by Italy in April 2023, albeit over privacy concerns, has precipitated studies in other European countries.[5] The new GPT-4 by OpenAI, touted to be revolutionary in how it can respond to both text and image commands, is available for a modest fee of US$20 a month to

1 "Pioneering Research on the Path to AGI" <https://openai.com/research/overview> (accessed 3 November 2023).

2 Krystal Hu, "ChatGPT Sets Record for Fastest-Growing User Base – Analyst Note", *Reuters* (2 February 2023) <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> (accessed 3 November 2023).

3 This should be contrasted with the term "AGI" which stands for Artificial General Intelligence, referring to "a theoretical type of artificial intelligence that possesses human-like cognitive abilities, such as the ability to learn, reason, solve problems, and communicate in natural language". See generally Gil Press, "Artificial General Intelligence (AGI) Is A Very Human Hallucination", *Forbes* (28 March 2023) <https://www.forbes.com/sites/gilpress/2023/03/28/artificial-general-intelligence-agi-is-a-very-human-hallucination/?sh=2f75b23364f2> (accessed 3 November 2023).

4 See, *eg*, *Authors Guild v OpenAI Inc*, Case 1:23-cv-08292 (SDNY, 2023); *Chabon v OpenAI, Inc*, Case 3:23-cv-04625 (ND Cal, 2023); *Silverman v OpenAI Inc*, Case 3:23-cv-03416 (ND Cal, 2023); *Tremblay v OpenAI Inc*, Case 3:23-cv-03223 (ND Cal, 2023); *Getty Images (US) Inc v Stability AI Inc*, Case 1:23-cv-00135 (D Del, 2023) and *Andersen v Stability AI Ltd*, Case 3:23-cv-00201 (ND Cal, 2023).

5 Supantha Mukherjee, Elvira Pollina & Rachel More, "Italy's ChatGPT Ban Attracts EU Privacy Regulators", *Reuters* (4 April 2023) <https://www.reuters.com/technology/germany-principle-could-block-chat-gpt-if-needed-data-protection-chief-2023-04-03/> (accessed 3 November 2023); Shiona McCallum, "ChatGPT Banned in Italy Over Privacy Concerns", *BBC* (1 April 2023) <https://www.bbc.com/news/technology-65139406> (accessed 3 November 2023).

ChatGPT Plus subscribers in the US. Not too long ago, many of us were obsessed with apps that can make us look like supermodels or superheroes, and today we are using a chatbot to help us write school essays and magazine articles, compose poems, and generate business proposals. The *Harvard Business Review* pointed out in February 2023 that while ChatGPT has created a global frenzy, "there are major practical, technical, and legal challenges to overcome before these tools can reach the scale, robustness, and reliability of an established search engine such as Google".[6] More recently, it highlighted a number of intractable intellectual property problems:[7]

> While it may seem like these new AI tools can conjure new material from the ether, that's not quite the case. Generative AI platforms are trained on data lakes and question snippets — billions of parameters that are constructed by software processing huge archives of images and text. The AI platforms recover patterns and relationships, which they then use to create rules, and then make judgments and predictions, when responding to a prompt.

3 Similarly, in a McKinsey & Company report released in June 2023 on the economic potential of generative AI, while it predicted that GAIAs could add trillions of dollars in value to the global economy, it warned that there are significant intellectual property risks in training data and model outputs, and that "organizations will need to understand what data went into training and how it's used in tool outputs".[8]

4 While the debate on whether autonomous AI-generated works deserve copyright protection appears to have momentarily taken a backseat, the present legal issues with AI systems that can produce essays or create realistic images and art from a description

6 Ege Gurdeniz & Kartik Hosanagar, "Generative AI Won't Revolutionize Search – Yet", *Harvard Business Review* (23 February 2023) <https://hbr.org/2023/02/generative-ai-wont-revolutionize-search-yet> (accessed 3 November 2023).

7 Gil Appel, Juliana Neelbauer & David A Schweidel, "Generative AI Has an Intellectual Property Problem", *Harvard Business Review* (7 April 2023) <https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem> (accessed 3 November 2023).

8 McKinsey & Company, *The Economic Potential of Generative AI: The Next Productivity Frontier* (June 2023) at p 49.

in natural language text prompts are very much occupying the centre stage in copyright law discussions. Nonetheless, with the US Copyright Office issuing registration guidelines in March 2023 that rejects the recognition of AI as author,[9] and the Supreme Court of the United Kingdom considering whether AI may be regarded as an "inventor" in patent law,[10] there is no doubt that AI authorship will remain on the agenda of policymakers in the years to come.

5 Singapore made headlines when it ambitiously revamped its Copyright Act[11] in 2021 that consolidated all previous amendments, rewrote the legislation in plain English and positioned the Act to be future-ready. The Singapore Copyright Act[12] was first enacted in 1987 and was largely based on the copyright regimes of the UK and Australia at that time. Major revisions to the Copyright Act were made in 1998, 1999 and 2004, which ensured that Singapore's copyright regime was aligned to international norms and bilateral treaties, and was relevant to content that was being created, distributed and consumed digitally. A significant public consultation exercise was carried out, which culminated in the introduction of the Copyright Act 2021[13] that came into force on 21 November 2021. Instead of adding significant amendments, this new legislation replaced the Copyright Act in its entirety. The wide-ranging reforms included the introduction of the moral right of attribution (or right to be identified), the recognition of an open-ended fair

9 The US Copyright Office has also released an authoritative statement of policy rejecting any registration of copyright for works created without any human contribution: US Copyright Office, Library of Congress, "Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence" (16 March 2023) <https://www.govinfo.gov/content/pkg/FR-2023-03-16/pdf/2023-05321.pdf> (accessed 3 November 2023). More recently, the District Court of the District of Columbia affirmed these principles — "Copyright has never stretched so far, however, as to protect works generated by new forms of technology operating absent any guiding human hand … Human authorship is a bedrock requirement of copyright": *Thaler v Perlmutter*, No 22-cv-1564, 2023 WL 5333236 at *4 (D.D.C, 18 August 2023).

10 *Thaler v Comptroller General of Patents, Trade Marks and Designs* [2021] EWCA Civ 1374.

11 Cap 63, 2006 Rev Ed.

12 Act 2 of 1987.

13 Act 22 of 2021.

use provision (modelled after the fair use provision in the US),[14] the inclusion of a computational data analysis (or text and data mining ("TDM")) exception, and a new class licensing scheme to regulate collective management organisations in Singapore.

6 The new Act was carefully calibrated to negotiate the complex relationships between protecting rights owners and enabling the public and other users to have access to these works in order to create new ones.[15] Significantly, by codifying an open-ended fair use provision akin to that in the US, works protected by copyright – which include music, videos, images and lyrics – may just be more readily available for transformative repurposing on social media platforms such as TikTok, Instagram and Facebook. However, at the time of public consultation in the mid-2010s, the GAIAs such as ChatGPT, DALL·E and DreamStudio were not even in the public consciousness.

7 This two-part article discusses how Singapore copyright law is poised to tackle two issues: (a) whether the use of copyright-protected works for machine learning ("input") and the works created from natural language commands ("output") are infringing copyright;[16] and (b) whether a TDM exception or fair use defence applies to such uses.[17]

II. Infringement – machine learning input and generated output

A. Machine learning input can potentially infringe copyright

8 In order for ChatGPT to respond to the questions or commands posed by human individuals, it needs to have access to millions or even billions of literary works – many of which are protected by copyright – in order to produce fully fleshed out answers and results based on digitally accessible text-based

14 Copyrights 17 USC (US) § 107.
15 See David Tan, "The Price of Generative AI Learning: Exceptions and Limitations under the New Singapore Copyright Act" (2023) 45 European Intellectual Property Review 400.
16 See paras 8–17 below.
17 This will be covered in Part 2 of the article.

information. Often referred to as the input of data for machine learning or machine training, an AI system is "fed" the relevant works in order for it to function effectively. OpenAI had previously revealed that in its earlier AI models, *eg*, GPT-1, it had accessed BookCorpus which had a collection of over 7,000 unique unpublished books. In 2020, while training GPT-3, the datasets came from two Internet-based books corpora amounting to 357,000 titles.[18] However, at the time of writing, the training datasets for GPT-4 have not been revealed. To date, most of the companies behind these impressive GAIAs have not disclosed the datasets they use for machine training. Nonetheless, in order for an AI application like DreamStudio to generate images based on text prompts, billions of text-and-image pairings have to be loaded into the computer memory, which are then encoded as an essential element of training the model. Stability AI, which owns Stable Diffusion and DreamStudio (a web-server-based AI image product), then adds visual noise to the encoded images to teach the model to generate output images that are consistent with a particular text description.

9 On 14 June 2023, the European Parliament of the European Union passed a draft law known as the Artificial Intelligence Act[19] (the "AI Act"), which would put new restrictions on what are seen as the technology's riskiest uses, and "would severely curtail uses of facial recognition software, while requiring makers of A.I. systems like the ChatGPT chatbot to disclose more about the data used to create their programs".[20] A final version of the law is not expected to be passed until later in 2023. It would seem that the issue took on new urgency for the 27-nation bloc after the

18 *Tremblay v OpenAI Inc*, Case 3:23-cv-03223 (ND Cal, 2023) at [28]–[35].

19 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 (2021), COM(2021) 206 final.

20 Adam Satariano, "Europeans Take a Major Step Toward Regulating A.I.", *The New York Times* (14 June 2023) <https://www.nytimes.com/2023/06/14/technology/europe-ai-regulation.html> (accessed 3 November 2023). See also 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 (2021), COM(2021) 206 final.

release of ChatGPT. According to an early report, under the latest version of Europe's bill:[21]

> … generative A.I. would face new transparency requirements. That includes publishing summaries of copyrighted material used for training the system, a proposal supported by the publishing industry but opposed by tech developers as technically infeasible.

While ChatGPT and other comparable text, image and video generative AI systems may not be "high-risk AI systems" as defined in Art 6 of the proposed AI Act, they would still be subject to the transparency obligations in Art 52.

10 The Singapore Copyright Act 2021[22] defines a "copy" of an authorial work as a reproduction of the work in any material form[23] and deems reproduction to have occurred if the work "is converted into … a digital or other electronic machine-readable form".[24] Furthermore, the making of a copy of a work that is temporary or is incidental to some other use of the work is to be treated as making a copy of the work.[25] Section 146 stipulates that copyright is infringed if a person, who is neither the copyright owner nor a licensee, does in Singapore, or authorises the doing in Singapore of, any act comprised in the copyright. Generally, copying entire or significant portions of works and permanently or temporarily storing them in computer memory constitutes *prima facie* infringement.

11 Presently, when inputting the images for machine learning, usually an algorithm will be scraping the Internet for content from various websites, invariably accessing content without permission and in violation of express prohibitions against such conduct contained in the terms of use of these websites. It is unlikely that all works used in GAIA training are open-access works or works in the public domain. Generally, in

21 Adam Satariano, "Europeans Take a Major Step Toward Regulating A.I.", *The New York Times* (14 June 2023) <https://www.nytimes.com/2023/06/14/technology/europe-ai-regulation.html> (accessed 3 November 2023).
22 2020 Rev Ed.
23 Copyright Act 2021 (2020 Rev Ed) s 41(1).
24 Copyright Act 2021 (2020 Rev Ed) s 41(2)(*f*).
25 Copyright Act 2021 (2020 Rev Ed) s 50(1).

the first stage of the data mining process (even if the AI system is not directly "fed" the relevant input), web robots may infringe the reproduction rights of the owners in the original literary, dramatic, musical and artistic (LDMA) works if such works are copied. Copying was established in the US decision of *Authors Guild Inc v Google Inc*,[26] where books were digitised in order to make them searchable, although this was consequently held to be fair use.[27] For instance, web robots that copy an artistic work such as paintings to gather information about the painting (*eg*, the number of brush strokes or the colour gradient) for further analysis, are infringing the reproduction rights to the paintings. Thus, an unauthorised reproduction of text and images would infringe copyright, regardless of whether the input for machine learning is consciously done by humans or automated by web robots.

12 Although the training data used for GAIAs has been kept a secret, more and more writers and visual artists are noticing similarities between their work and the output from these systems. It is virtually impossible to prove that a particular work has been used as a training *input* for a GAIA unless the *output* exhibits substantial similarity to the original work. In the *Andersen v Stability AI*[28] claim in California, the motions to dismiss filed by the defendants in April 2023 all indicate, *inter alia*, that the plaintiffs failed to identify a single actual output image that allegedly infringes any of the copyrighted works. On the contrary, in *Getty Images (US) Inc v Stability AI Inc*[29] ("*Getty Images v Stability AI*"), where Getty Images filed a lawsuit against Stability AI in the US for copying over 12 million photographs from its collection, there was evidence of output images that contained Getty Images' watermark. Nonetheless this only indicates that a particular work was used for AI training, and does not provide sufficient evidentiary support for a claim of wholesale copying of millions of works. More recently in September 2023, authors Michael Chabon (a Pulitzer Prize recipient), David Henry

26 804 F 3d 202 (2nd Cir, 2015).
27 *Authors Guild v Google Inc*, 804 F 3d 202 at 207 (2nd Cir, 2015).
28 *Andersen v Stability AI Ltd*, Case 3:23-cv-00201 (ND Cal, 2023).
29 Case 1:23-cv-00135 (D Del, 2023).

Hwang (a Tony Award-winning playwright) and several others filed a claim against OpenAI for infringement of their literary works which they alleged are used as training datasets to power its ChatGPT product. It would appear that in these individual scenarios, the output provided compelling evidence that each of the specific author's works was used as input.[30]

13 There are two additional hurdles. First, in the US, there is a mandatory requirement of registration of copyright with the US Copyright Office before one can commence a claim for infringement.[31] Alleging wholesale copying of thousands or millions of works in the input would not clear this hurdle unless each and every work has been registered. Second, some GAIAs do not store the copyrighted works at all, but rather mathematical representations of patterns collected from these images. For instance, Stability AI utilises datasets prepared by a third party – LAION (a German entity) which created datasets of image-text pairs by scraping links to billions of works from various websites – encodes the images and adds visual noise to teach the AI how to decode the image and remove noise in order to generate a final output image. Claimants would have to prove that the defendant, *eg*, Stability AI, has made at least a temporary reproduction of the relevant work as input in the AI training. Under s 41 of the Singapore Copyright Act 2021,[32] the conversion of authorial works into a machine-readable format would be deemed to be a reproduction.

14 In July 2023, the Authors Guild sent an open letter to the leaders of some of the world's biggest generative AI companies. Signed by more than 9,000 writers, including prominent authors such as George Saunders and Margaret Atwood, it asked the likes of Alphabet, OpenAI, Meta, and Microsoft "to obtain consent, credit, and fairly compensate writers for the use of copyrighted

30 *Chabon v OpenAI Inc*, Case 3:23-cv-04625 (ND Cal, 2023) at [49]–[53].
31 *Fourth Estate Public Benefit Corp v Wall-Street.com LLC* 139 S Ct 881 (2019) at 885–888; Copyrights 17 USC (US) § 411(a).
32 2020 Rev Ed.

materials in training AI".[33] More recently, perhaps an implicit acknowledgment that the input of copyright-protected works for AI learning would be infringing, Adobe announced that it plans to pay content creators for using their works in its new AI tool – Firefly – and that it will also allow creators the choice to not let their works be used to train the AI.[34]

B. Generated output can also infringe copyright

15 When ChatGPT or DreamStudio generates text or images based on the user's questions or commands, the output can also infringe copyright in a source text or image if it is substantially similar to the original. In theory, ChatGPT's output, like other LLMs, will be generated based on patterns and connections drawn from the training data. If asked to generate an essay on the position of copyright fair use in Singapore, ChatGPT is unlikely to paraphrase all the sentences from its training dataset of literary works, and will invariably reproduce significant amount of text *verbatim* from its sources (which may include academic articles, court judgments and online commentaries). In the *Getty Images v Stability AI* lawsuit, the claim identified some of the output delivered by Stability AI in the DreamStudio application to include a modified or distorted version of a Getty Images watermark, underscoring the clear link between the copyrighted images and the final product. In such circumstances, this would be another instance of copyright infringement. However, in the class action lawsuit by artists that included Sarah Andersen, the motions to dismiss filed by the defendants' lawyers pointed out that the plaintiffs failed to clearly identify which particular output was substantially similar, and hence infringing, to a specific input.

16 One should further note that copyright does not protect the style of an artist, no matter how distinctive – this includes a

33 Will Bedingfield, "The Generative AI Battle Has a Fundamental Flaw", *Wired* (25 July 2023) <https://www.wired.co.uk/artificial-intelligence-copyright-law?verso=true> (accessed 3 November 2023).

34 Krist Boo, "Adobe to Pay AI-Content Creators in Groundbreaking Move", *The Straits Times* (21 March 2013) <https://www.straitstimes.com/business/adobe-to-pay-ai-content-creators-in-groundbreaking-move> (accessed 3 November 2023).

painting-style (like Picasso's distinctive cubist style or Warhol's silkscreen treatments of photographs), writing-style and singing-style. The artistic style of an author in copyright law is generally considered an "idea" in the well-established idea-expression dichotomy, which has been codified in the US, and also adopted by the Singapore Court of Appeal.[35] In the same way that we can freely paint and sell a scenery of the Singapore Botanic Gardens in a Monet impressionist-style (assuming that Claude Monet's paintings are still protected by copyright), it is not copyright infringement if DALL·E, in response to a prompt "Singapore Botanic Gardens in the style of Monet" generates a particular image that evokes Monet's *Bridge Over A Pond Of Water Lilies*. The assessment of infringing outputs is a fact-intensive inquiry where one would need to prove substantial similarity between the output text/image and the expression protected in the original input text/image, and one should not presume that if the storage of *input* for generative AI learning was infringing, the *output* would necessarily be infringing.

17 In summary, it is difficult to prove wholesale copying of millions of works as the various GAIAs do not disclose the training datasets, and one would have to proceed on a classic substantial similarity analysis in respect of *each* output text/image *vis-à-vis* the original work. Perhaps realising the enormity of the task before them should they resort to litigation, media companies such as CNN, The New York Times and Reuters, have deployed technological defensive measures such as injecting code into their websites that blocks OpenAI's web crawler, GPTBot, from scanning their platforms for content.[36]

35 *Global Yellow Pages Ltd v Promedia Directories Pte Ltd* [2017] 2 SLR 185 at [15].

36 Oliver Darcy, "Disney, The New York Times and CNN Are Among a Dozen Major Media Companies Blocking Access to ChatGPT as They Wage a Cold War on A.I.", *CNN* (28 August 2023) <https://edition.cnn.com/2023/08/28/media/media-companies-blocking-chatgpt-reliable-sources/index.html>.