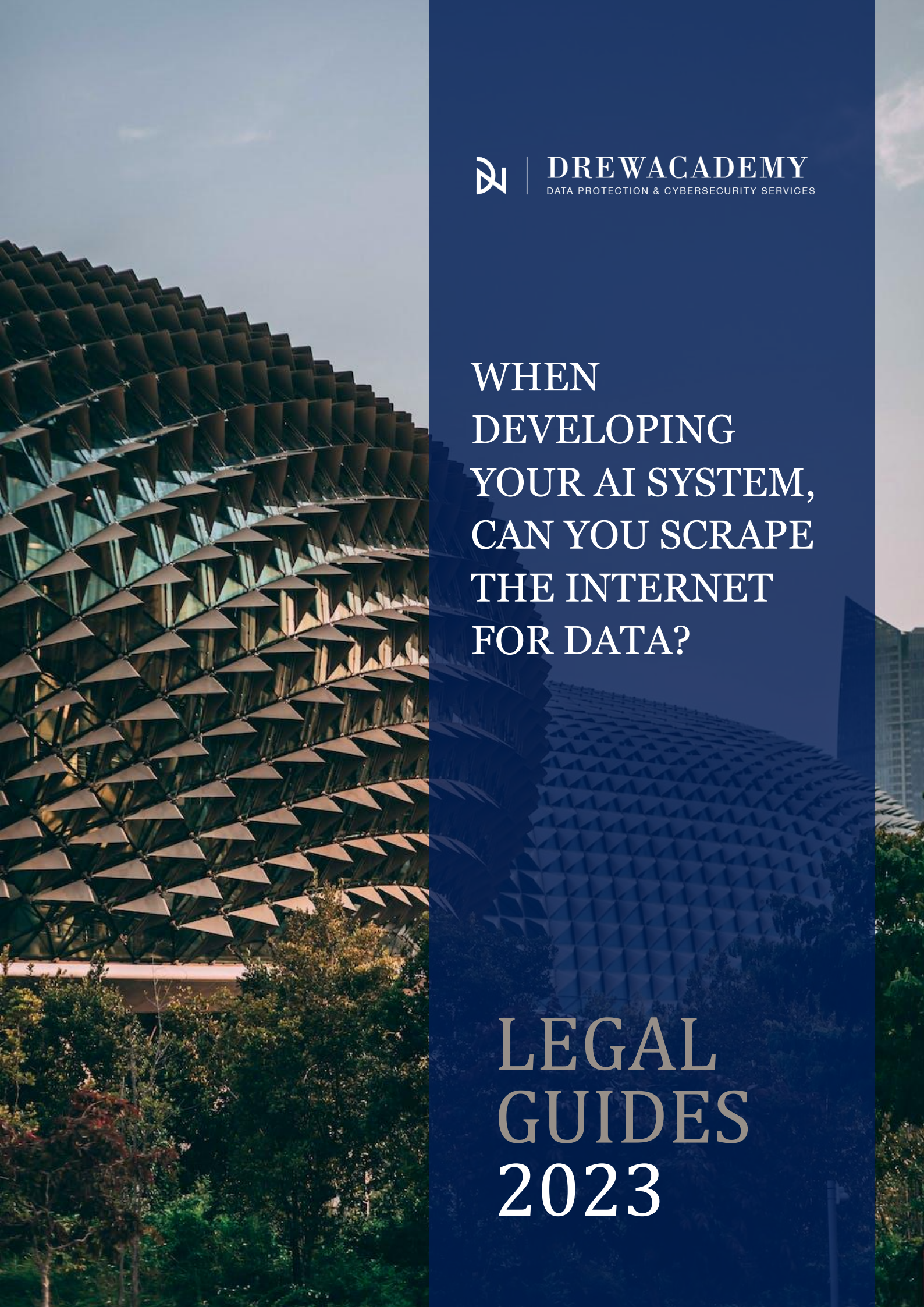

DREWACADEMY
DATA PROTECTION & CYBERSECURITY SERVICES

WHEN DEVELOPING YOUR AI SYSTEM, CAN YOU SCRAPE THE INTERNET FOR DATA?

LEGAL GUIDES 2023

CONTENTS

Developments as of 11 April 2023

DREWACADEMY
DATA PROTECTION & CYBERSECURITY SERVICES

WHEN DEVELOPING YOUR AI SYSTEM, CAN YOU SCRAPE THE INTERNET FOR DATA?

In the last 6 months, lawsuits against Microsoft and Stability AI in the US and UK have made headlines for the potential ramifications they may have on the artificial intelligence ("AI") space. At the heart of these two cases is the question of whether AI systems that scrape the internet for data infringe the copyright of individuals and corporations.

Code generators

In November 2022, Matthew Butterick filed a lawsuit against Microsoft's subsidiary, GitHub, and its business partner, OpenAI, over the companies' AI-powered coding assistant, GitHub Copilot.[1] Officially launched as a paid subscription service in June 2022, Copilot uses OpenAI's technology to generate and suggest lines of code directly within the programmer's code editor. The programmer simply needs to describe the function they wish to code in natural language and Copilot will generate an AI-suggested block of code that the programmer can then use or adapt.[2] Copilot's generative output ability has been trained based on "billions of lines of code" from publicly available sources, including code in public repositories on GitHub.[3]

However, the lawsuit claims that the data and code used to train Copilot infringes on copyright and open-source code licences. While open-source software is typically offered for users to use and modify freely, the use of such software often requires compliance with terms of the licence which include attribution and inclusion of a copyright notice. Copilot, however, has been found to copy and produce long sections of licenced code without providing any credit.[4] This lawsuit is still in its early stages but the core issue before the courts is likely to be whether the use and copying of these codes falls under the doctrine of fair use under US copyright law.

Image generators

In January 2023, Getty Images, a renowned stock photo provider, took the first step in commencing legal proceedings in the UK against Stability AI, the company behind the AI image generator Stable Diffusion.[5] Getty Images has also filed a similar lawsuit in the US in February 2023.[6] Getty Images alleges that Stability AI unlawfully copied and processed images and associated metadata it owns without a licence in order to develop Stable Diffusion. Stable Diffusion is trained by being fed images and then reconstructing the images with maximum accuracy. It is then able to generate new images from a text prompt. Stable Diffusion has stated that the dataset used in training the AI is from LAION, a non-profit, that does a "general crawl of the internet"[7] and an independent analysis of this dataset found that Getty Images and other stock image sites constituted a large portion of its content[8].

What is Singapore's position on the use of data to train AI systems?

While such questions have not been raised for consideration in the Singapore courts yet, Parliament has made some legislative changes to keep pace with technological changes. In 2019, the Ministry of Law and the Intellectual Property Office of Singapore released the Singapore Copyright Review Report ("Copyright Review Report").[9] The report noted that text and data mining and its applications were crucial to driving economic growth and innovation in the digital economy.[10] While the doctrine of fair use was and continues to be available under the copyright regime in Singapore, the report proposed that the better option was to introduce a specific exception to copyright infringement for text and data

[1] Doe v. GitHub Inc, U.S. District Court for the Northern District of California, No. 4:22-cv-06823
[2] https://github.com/features/copilot.
[3] *Ibid*.
[4] James Vincent (2022), *The lawsuit that could rewrite the rules of AI copyright*.
[5] Getty Images (2023), *Getty Images Statement*.
[6] Getty Images (US) Inc v. Stability AI Inc, U.S. District Court for the District of Delaware, No. 1:23-cv-00135
[7] https://stability.ai/faq; https://stablediffusionweb.com/ (see Frequently Asked Questions)
[8] See "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator" by Baio (2022).
[9] https://www.mlaw.gov.sg/files/news/press-releases/2019/01/Annex%20A%20-%20Copyright%20Review%20Report%2016%20Jan%202019.pdf.
[10] Paragraph 2.8.4 of the Copyright Review Report.

analysis.[11] The report also clarified that the exception will apply to both non-commercial and commercial uses.[12] Thus, in the Copyright Act 2021, the computational data analysis exception was introduced under sections 243 and 244 as a permitted use of copyrighted works.

Section 243 of the Copyright Act 2021 defines "computational data analysis" as including, in relation to a work or recording of a protected performance, (a) using a computer program to identify, extract and analyse information or data from a work or recording and (b) using the work or recording as an example of a type of information or data to improve the functioning of a computer program in relation to that type of information or data.

Section 244 then provides that it is a permitted use of a copyrighted work where a copy of the work is made for the purpose of computational data analysis or preparing the work for computational data analysis. However, the person who copies the work must not use the copy for any other purpose and must have lawful access to the work from which the copy is made.[13] Additionally, the person must not supply the copy of the work to any other person other than for the purpose of verifying results of the computational data analysis or collaborative research relating to the purpose of the computational data analysis.[14]

At first instance it would therefore appear that if a claim similar to that in Github and Stability Diffusion were to arise in Singapore, it would be easily disposed of under section 244 of the Copyright Act 2021. However, it is arguable that the answer is not so clear. A distinction may be drawn between AI that carries out an analytical function and a generative function. The definition provided under section 243 of the CA makes reference to analysing and extracting information or using the work as an example to improve the functioning of the AI (e.g. using images to train a computer program to recognise images)[15]. Therefore, the Copyright Act 2021 seems to permit the use of copyrighted work in relation to training AI that have an analytical function as opposed to a generative one. The use of the computational data analysis exception in relation to output generating AI may also be contradictory to the policy objective stated in the Copyright Review Report. The report stated that the purpose of the exception was to "analyse data and not to consume what copyright seeks to protect".[16]

However, the definition of "computational data analysis" provided under section 243 of the CA is open-ended and it may be open to the courts to deem the training of generative AI with copyright protected works as permissible. This may boil down to a factual analysis of how the AI functions. As copyright seeks to protect the expression of ideas only, if the AI is found to only analyse and extract information from the expression rather than copy the expression itself to generate output, this may fall under the computational data analysis exception.

The UK currently has a similar provision under the Copyright Design and Patents Act 1988 ("CDPA") with respect to permissible copying of work but this only applies to non-commercial research.[17] However, the UK government has also acknowledged the importance of text and data mining to support innovation and intends to introduce a new copyright and database right exception to allow text and data mining for any purpose.[18]

[11] Paragraph 2.8.5 of the Copyright Review Report.
[12] Paragraph 2.8.6 of the Copyright Review Report.
[13] Section 244(2) of the CA.
[14] Section 244(2)(c) of the CA.
[15] Illustration to section 243 of the CA.
[16] Paragraph 2.8.6 of the Copyright Review Report.
[17] Section 29A of the CDPA.
[18] See "*Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation*" by UK Intellectual Property Office (2022) at paragraph 58.

Conclusion

The legality of using data, especially data sourced from third-party sources or the Internet, to train AI systems, is still a developing area of law. The lawsuits in other jurisdictions will have a bearing on how the data can be used, and we are watching this space closely. In the meantime, organisations should bear in mind the conditions for lawful computational data analysis under section 244 of the Copyright Act 2021 to avoid infringing copyright.

Drew Academy wishes to acknowledge our Associate Julian Liaw for assisting in the preparation of this article.

DREW DATA PROTECTION & CYBERSECURITY ACADEMY

Drew Data Protection & Cybersecurity Academy (Drew Academy) was established in 2020 by Drew & Napier to help our clients build their capabilities and develop and implement organisational strategies, structures, policies and processes to meet their legal, regulatory and compliance obligations. Drew Academy offers a range of courses in areas such as data protection, cybersecurity, data governance and in-house commercial practice. A particular focus for us is the delivery of workplace learning solutions and development of customised training courses. We also offer outsourced DPO services and data protection consulting services through our experienced team of practitioners.

Drew Academy is helmed by Lim Chong Kin and David N. Alfred. Our course leaders are experienced in various aspects of data and cyber governance, data protection, cybersecurity engineering and in-house commercial practice.

ARTIFICIAL INTELLIGENCE AND DIGITAL TRUST

Drew & Napier's Artificial Intelligence (AI) and Digital Trust practice brings together its expertise across several technology-related domains and in fields as diverse as data protection, cybersecurity, healthcare, Fintech, intellectual property and competition law (to name a few) to advise clients on the full range of legal issues relating to AI and Digital Trust. In addition to advising on commercial, regulatory and international / cross-border issues, our advice extends into areas such as governance and ethics as we seek to enable our clients to navigate areas where laws and legal principles are still emerging.

Working together with the Drew Academy, we provide solutions that reflect our deep understanding of underlying technologies, the risks and uncertainties involved and practical business considerations. Internationally, there is a growing consensus on AI governance.

For more information on our experience, please contact:

Lim Chong Kin
Managing Director, Corporate & Finance;
Co-Head, Data Protection,
Privacy & Cybersecurity Practice;
Co-Head, Drew Data Protection &
Cybersecurity Academy

T: +65 6531 4110
E: chongkin.lim@drewnapier.com

Benjamin Gaw
Director, Corporate and
Merger & Acquisitions;
Head, Healthcare & Life Sciences
(Corporate & Regulatory)

T: +65 6531 2393
E: benjamin.gaw@drewnapier.com

David N. Alfred
Director, Corporate & Finance;
Co-Head, Data Protection,
Privacy & Cybersecurity Practice;
Co-Head and Programme Director,
Drew Data Protection &
Cybersecurity Academy

T: +65 6531 2342
E: david.alfred@drewnapier.com

Cheryl Seah
Director, Corporate & Finance

T: +65 6531 4167
E: cheryl.seah@drewnapier.com

DREWACADEMY
DATA PROTECTION & CYBERSECURITY SERVICES

10 Collyer Quay
10th Floor Ocean Financial Centre
Singapore 049315

www.drewnapier.com/Academy

T: +65 6531 4152
F: +65 6535 4864
E: academy@drewnapier.com

In association with

DREW & NAPIER