Should Hong Kong introduce an explicit copyright exemption for synthetic data to immunize users and creators of synthetic data?

Response to the Public Consultation on Copyright and Artificial Intelligence (AI) by

- Martin Seng, JD Graduate, City University of Hong Kong
- Ronald Yu, Director, Makebell Ltd.; Visiting Fellow City University of Hong Kong

Artificial intelligence, in particular large language models (LLMs), require vast amounts of data. While this data can originate from existing sources, data is increasingly being generated synthetically as, in many cases, there is insufficient data to train LLMs for specific applications, particularly when something new emerges (e.g. new laws or regulations).

As LLMs require large amounts of data for training, this hampers their ability to deal with new developments as sufficient real-world data will not immediately exist.

What is Synthetic Data?

Synthetic data, which may be defined as artificially generated data using real, organically generated data as input, can overcome such limitations.

Synthetic data has the same statistical properties as real data and enables faster development of LLMs and their associated applications as it can be generated quickly and efficiently unlike real-world data which can take months or even years to generate in sufficient quantities.

In addition, synthetic data can also be more cost-effective than collecting and processing large amounts of real-world data[1] and creates technological, economic and ethical opportunities, including the potential to:

- Improve accuracy by mitigating the unreliability of human-made data, which is typically gathered by scraping the erratic web that is the internet
- Mitigate or even remove biases and imbalances in existing, human-made data[2]
- Reduce the costs and obstacles at all stages of the data value chain, which may help by lowering costs of developing data and removing data barriers to entry in relevant markets, characterized by network effects[3] and
- Offer opportunities for companies specializing in providing synthetic datasets, either from a pre-existing proprietary database or by creating "bespoke" synthetic data generated on demand for specific customers[4] who are starting to run out of easily accessible, reliable and high-quality real-world data sources to continue training more advanced AI models, or have become concerned about potential intellectual property (IP) issues with real-world data.

Issues related to Synthetic Data

However, creating synthetic data may itself require training on real-world data samples that may contain copyrighted works, as well as materials displaying trademarks, data compilations which may be protected by *sui*

[1] Kaled El Emam (2020) Accelerating AI with Synthetic Data: Generating Data for AI Projects
[2] Emiliano De Cristofaro (2024), Synthetic Data: Methods, Use Cases, and Risks
[3] Gareth Kristensen, Angela Dunning, Gaia Shen, Prudence Buckland, Jan-Frederik Keustermans & Alix Anciaux (2023), Training AI models on Synthetic Data: No silver bullet for IP infringement risk in the context of training AI systems (Part 1 of 4)
[4] For example, *see* https://scale.com/; https://gretel.ai/; and https://mostly.ai/.

generis database rights in other jurisdictions or personal data, potentially triggering IP and data privacy issues.[5]

As Hong Kong does not have an explicit exemption for text and data mining (TDM) that would, for example, obviate the need for developers and users of data to obtain permission from multiple rights holders in order to access the data and also avoid extra licence fees,[6] the resulting ambiguity regarding IP infringement or other risks for users or providers of synthetic data may worry creators and users of synthetic data, in turn potentially hampering AI development in Hong Kong.

For those reasons, Hong Kong should consider a TDM exemption that removes the aforementioned ambiguity regarding IP infringement for AI developers as well as users and creators of synthetic data.

[5] Gareth Kristensen, Angela Dunning, Gaia Shen, Prudence Buckland, Jan-Frederik Keustermans & Alix Anciaux (2023), Training AI models on Synthetic Data: No silver bullet for IP infringement risk in the context of training AI systems (Part 2 of 4)

[6] Oliver Bray (2022), UK announces new copyright exemption for text and data mining to promote AI development