KOREA COPYRIGHT COMMISSION

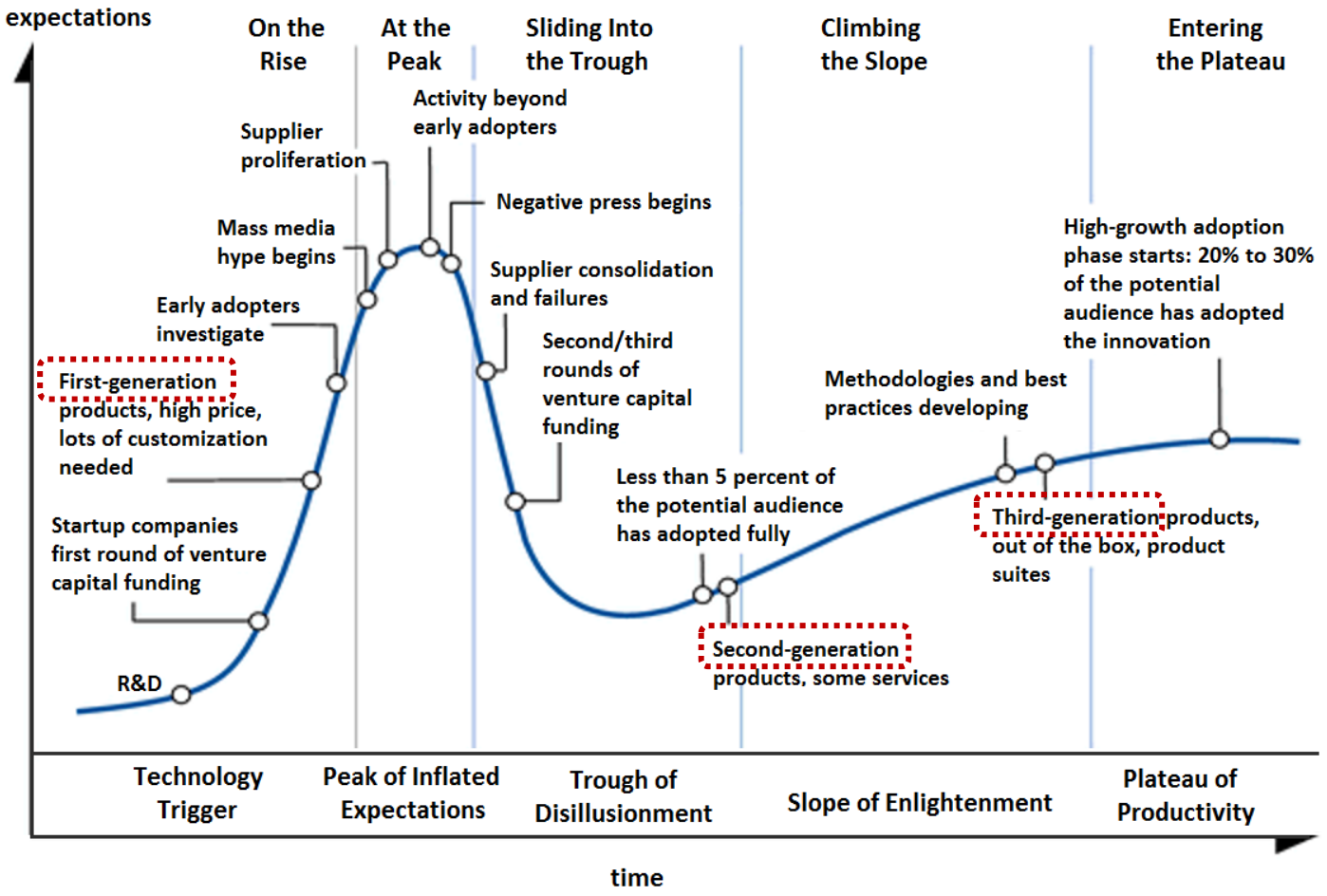
Public Domain & Open Source SW
International Conference 2024

Open Culture in the Era of AI:
New Horizons and Opportunities

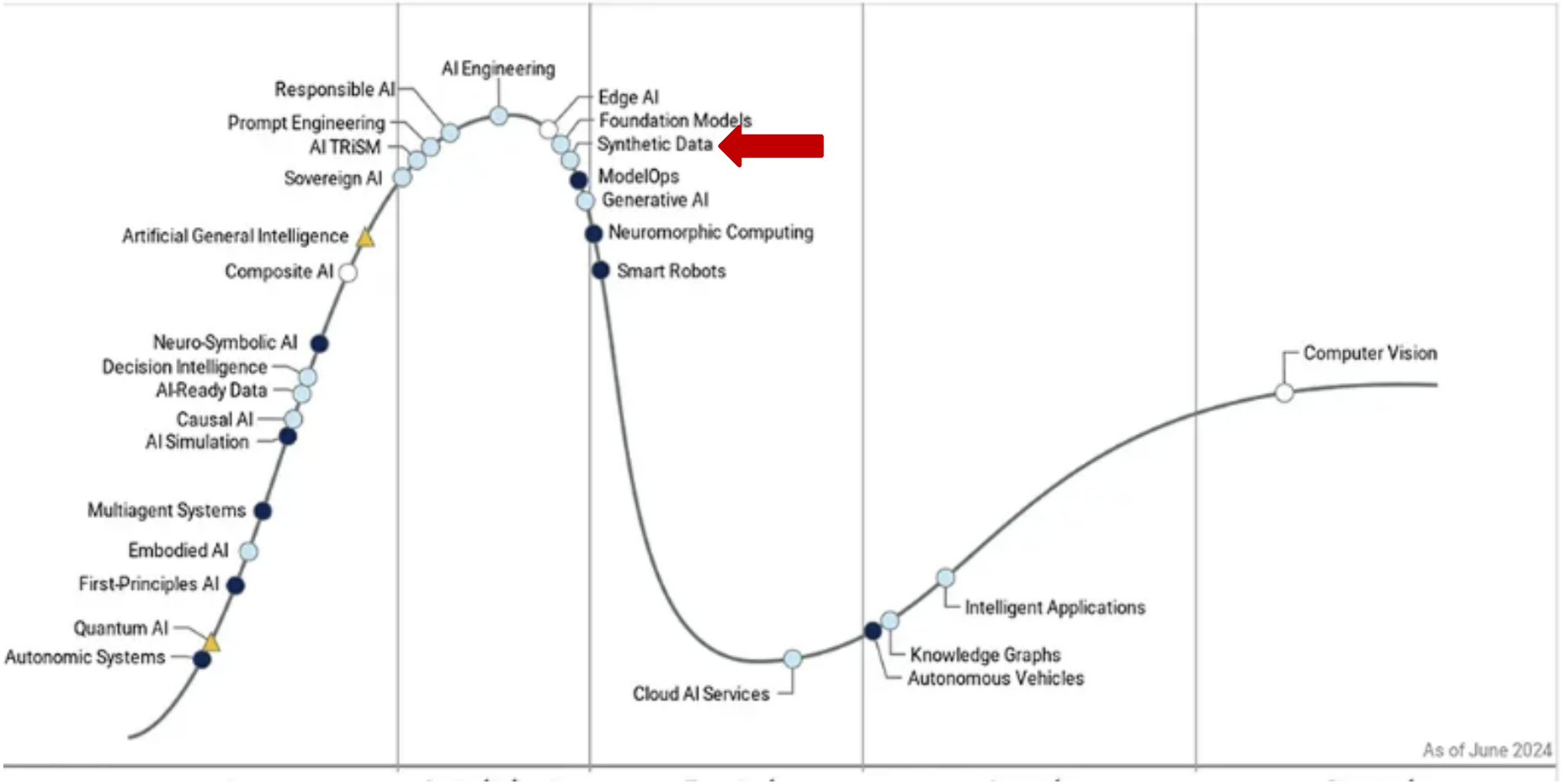
Myuhng-Joo Kim

Vice chair, Korea Copyright Commission
Professor, Seoul Women's University
President, International Association for AI & Ethics
Member, OECD GPAI (Global Partnership on AI)

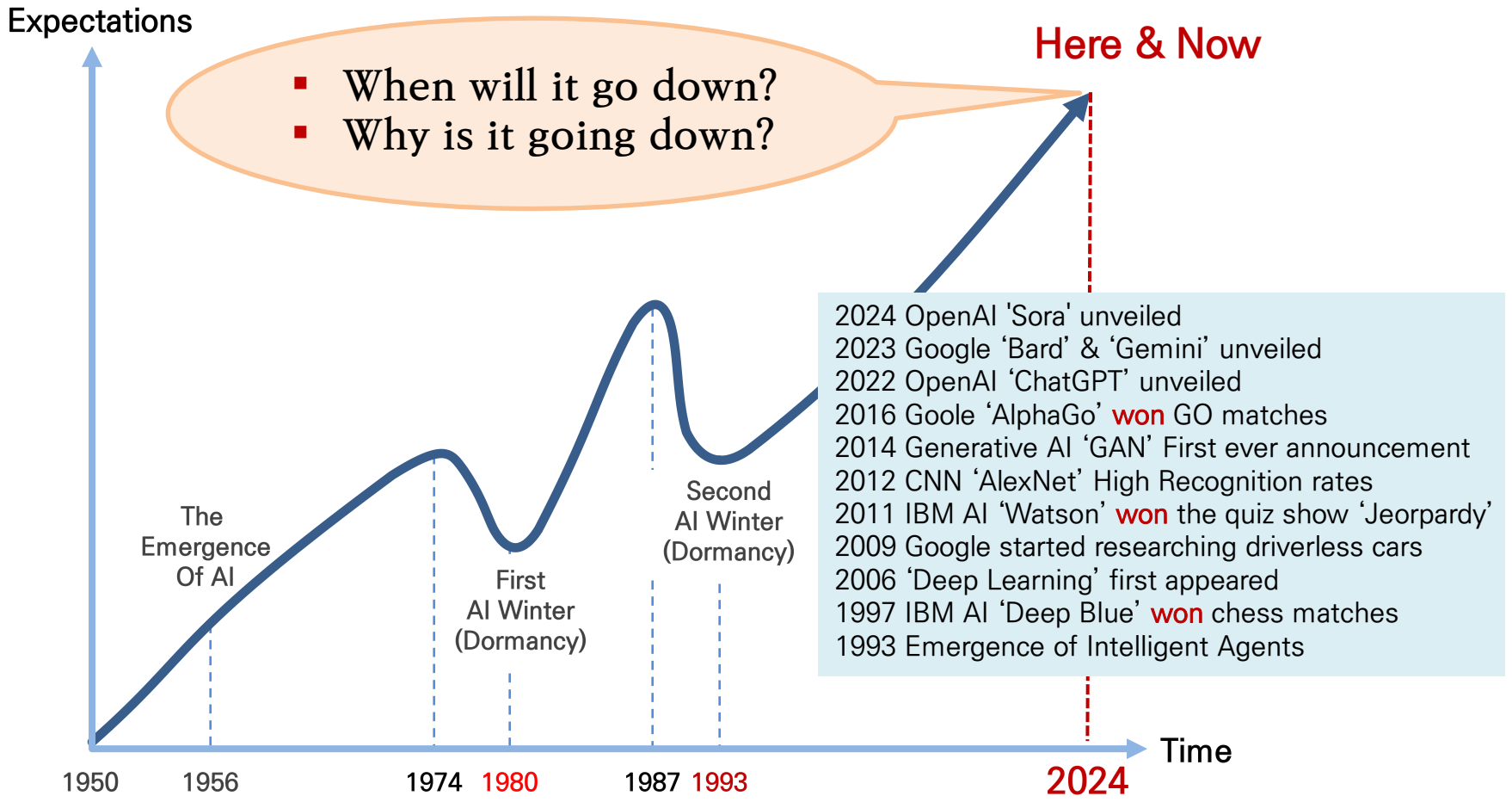
Hype Cycle (Gartner group, Sep., 2005)



Hype Cycle for AI, 2024

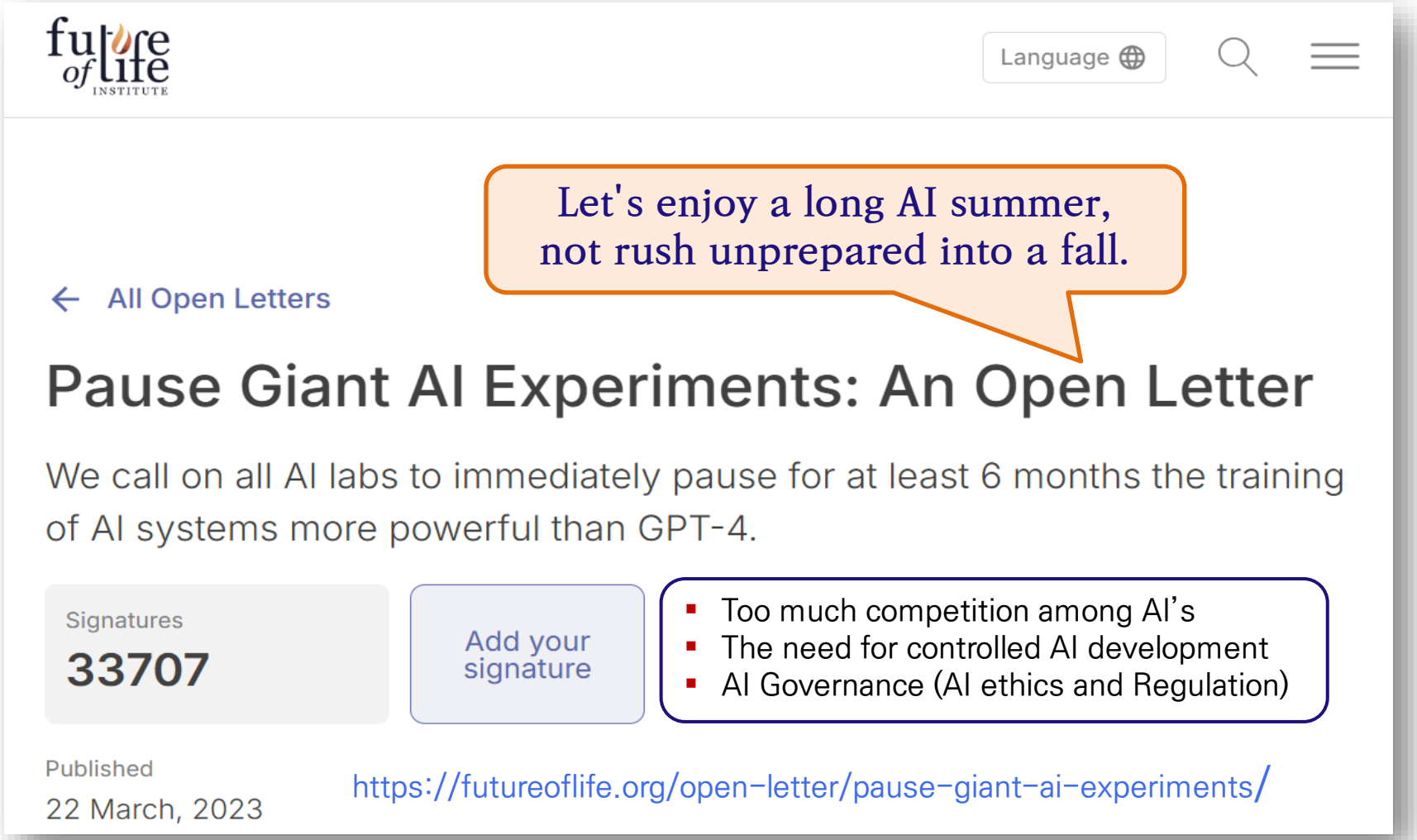


70 years of Artificial Intelligence



https://www.actuaries.digital/2018/09/05/history-of-ai-winters/

An Open Letter from FLI

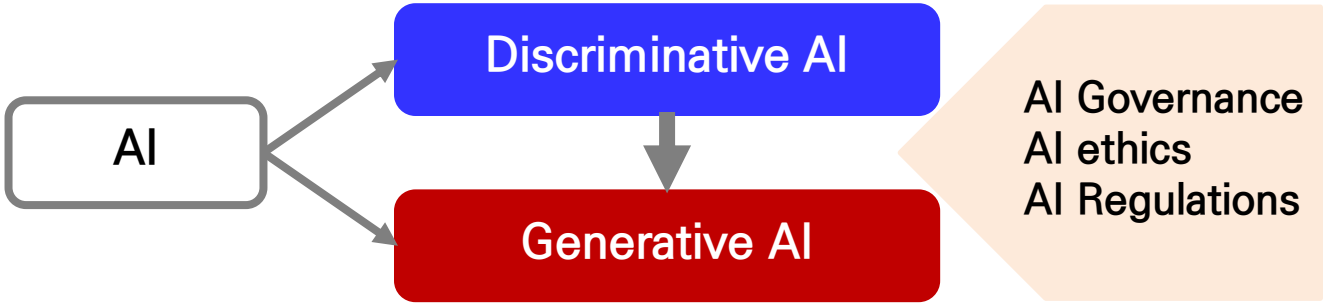


The Legal Definition on AI

European Union's AI Act 〈13th March 2024〉

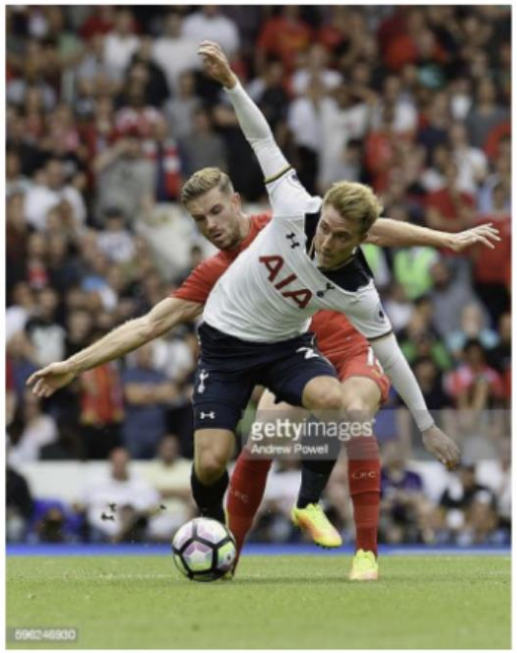
- 'AI system' means a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

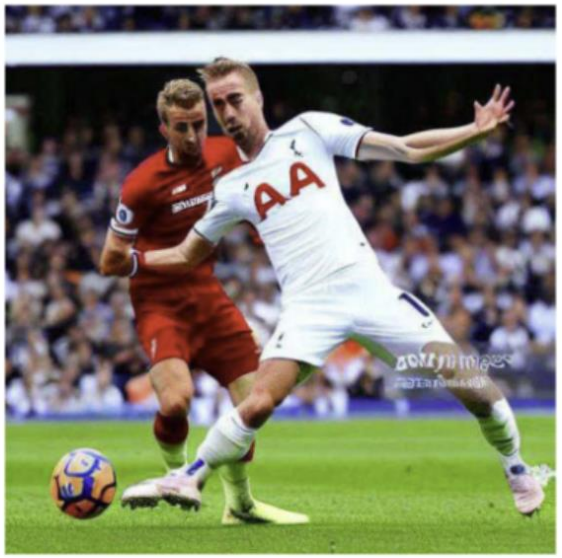
(Article 3 Definitions)



Copyright infringement lawsuits

Getty Images files litigation against AI image generation biz for IP infringement





- 17th Jan 2023
- It was argued that UK start-up, Stability AI, utilized 12 million images owned by Getty Images during the training of its AI image generator, Stable Diffusion
- Filed litigation for IP infringement at court in Delaware, US, and High Court, London

Source: Image from Getty Images and the image by Stability AI under legal contention

Copyright royalty lawsuits

NYT sues Open AI, demanding royalty payment for news articles

- 29th Dec 2023
- NYT: As biggest biz in news industry, had led the way for paid online content
- When negotiations broke down, filed litigation against Open AI (developer of ChatGPT) and Microsoft for ① Copyright infringement of its news articles ② Libel as result of fake news generated by AI
- Using news articles for training generative AI model is copyright infringement, not fair use

ChatGPT now pays WSJ for its news articles… Paid USD 250 million to News Corp

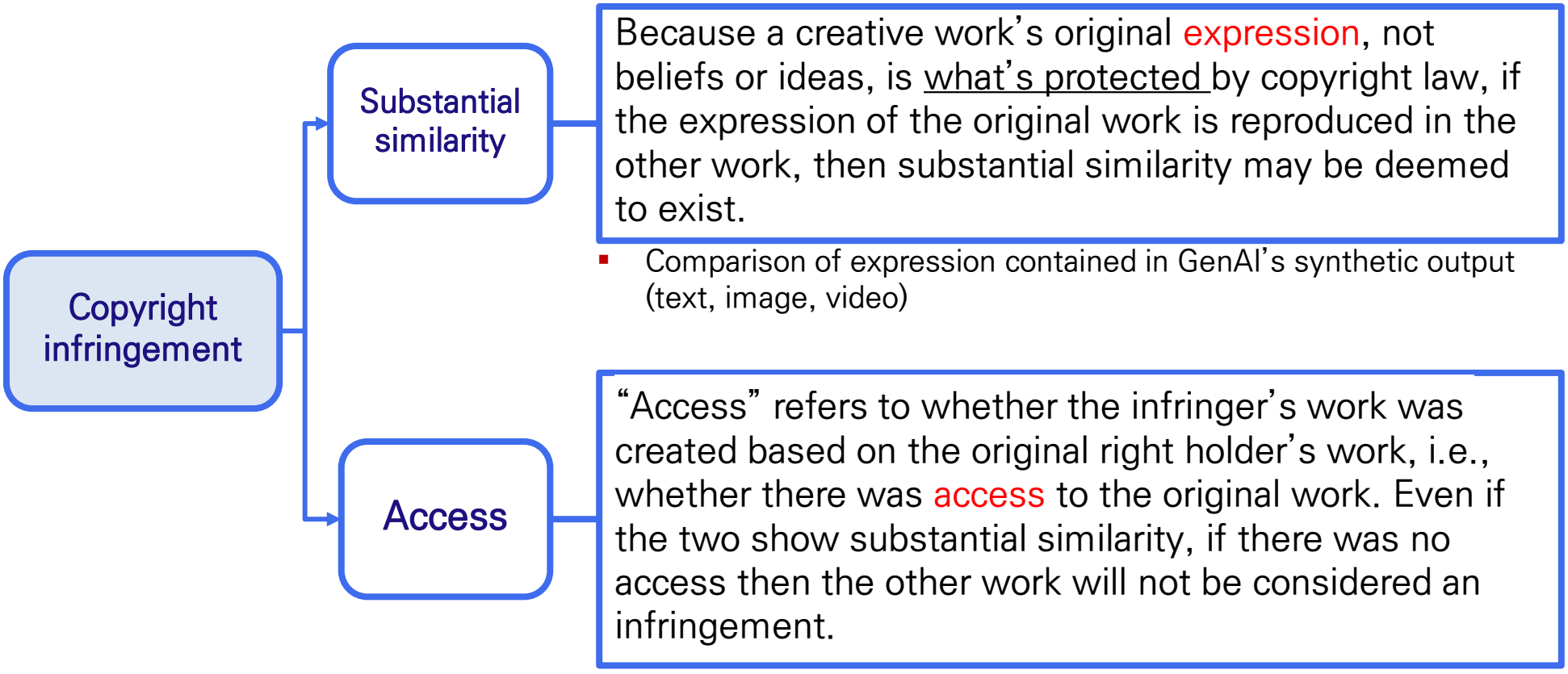
Joongang Daily, 23rd May 2024

- 23rd May 2023
- Open AI signed deal with News Corp (owner of NYT, NYP, UK's The Times & The Sun, Australia's Sky News)
- Will be paid USD 250 mil (KRW 340 bil) during next 5 years

Announcement of Naver's ChatGPT leads to row over news royalty

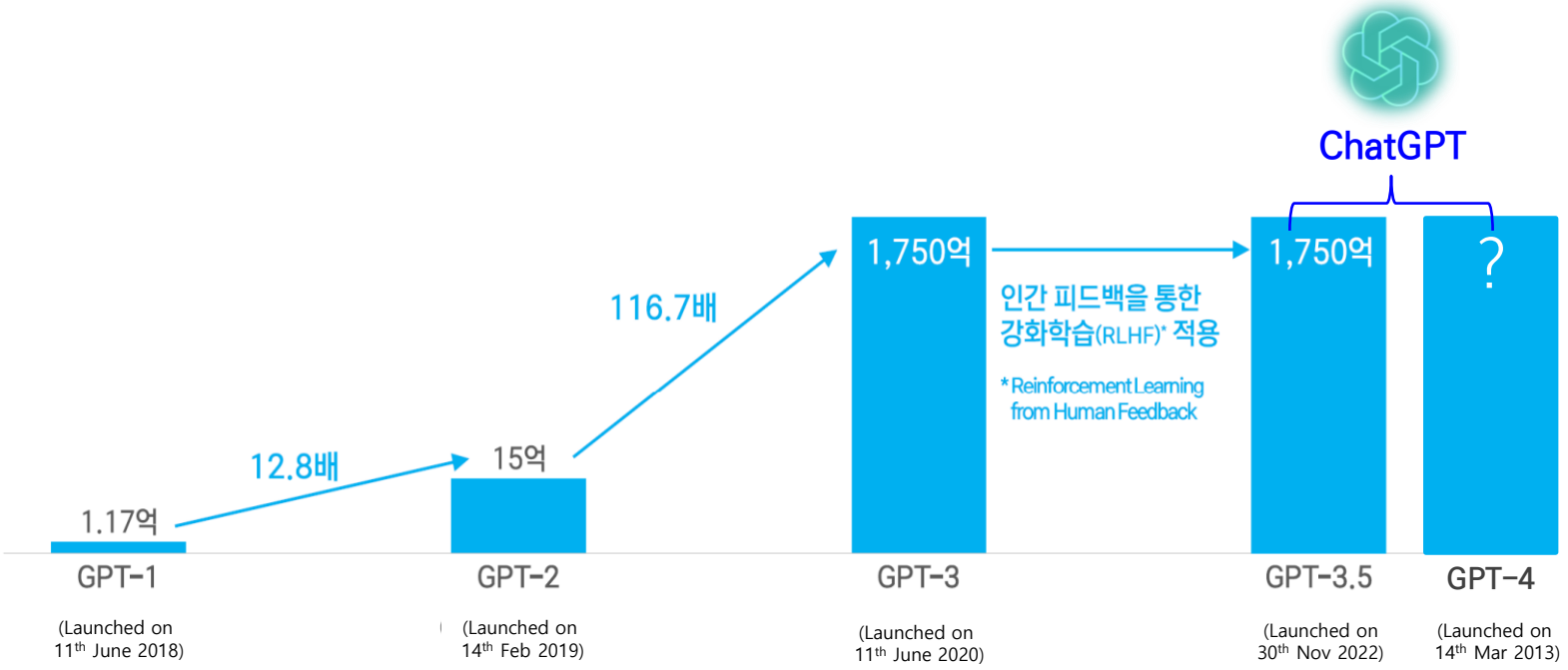
- 26th Aug 2023
- Korea Newspaper Assoc, Korea Press Foundation vs Naver

Whether GenAI infringes copyright



Do Language Models Plagiarize?

- Lee, J., Le, T., Chen, J., & Lee, D. (2023). Do Language Models Plagiarize? In *ACM Web Conference 2023*
- 210,000 articles generated by GPT-2
  vs 8,000,000 articles used as training data of GPT-2



Do Language Models Plagiarize?

Verbatim

Paraphrase

Idea

Type	Machine-Written Text	Training Text
Verbatim	*** is the second amendment columnist for Breitbart news and host of bullets with ***, a Breitbart news podcast. [...] (***Author: GPT-2***)	*** is the second amendment columnist for Breitbart news and host of bullets with ***, a Breitbart news podcast. [...]
Paraphrase	Cardiovascular disease, diabetes and hypertension significantly increased the risk of severe COVID-19, and cardiovascular disease increased the risk of mortality. (***Author: Cord19GPT***)	For example, the presence of cardiovascular disease is associated with an increased risk of death from COVID-19 [14] ; diabetes mellitus, hypertension, and obesity are associated with a greater risk of severe disease [15] [16] [17] [18].
Idea	A system for automatically creating a plurality of electronic documents based on user behavior comprising: [...] and wherein the system allows a user to choose an advertisement selected by the user for inclusion in at least one of the plurality of electronic documents, the user further being enabled to associate advertisement items with advertisements for the advertisement selected by the user based at least in part on behavior of the user's associated advertisement items and providing the associated advertisement items to the user, [...] . (***Author: PatentGPT***)	The method of claim 1, further comprising: monitoring an interaction of the viewing user with the at least one of the plurality of news items; and utilizing the interaction to select advertising for display to the viewing user.

Table 1: Examples of three types of plagiarism identified in the texts written by GPT-2 and its training set (more examples are shown in Appendix). Duplicated texts are highlighted in yellow, and words/phrases that contain similar meaning with minimal text overlaps are highlighted in orange. [...] indicates the texts omitted for brevity. Personally identifiable information (PII) was masked as *.**

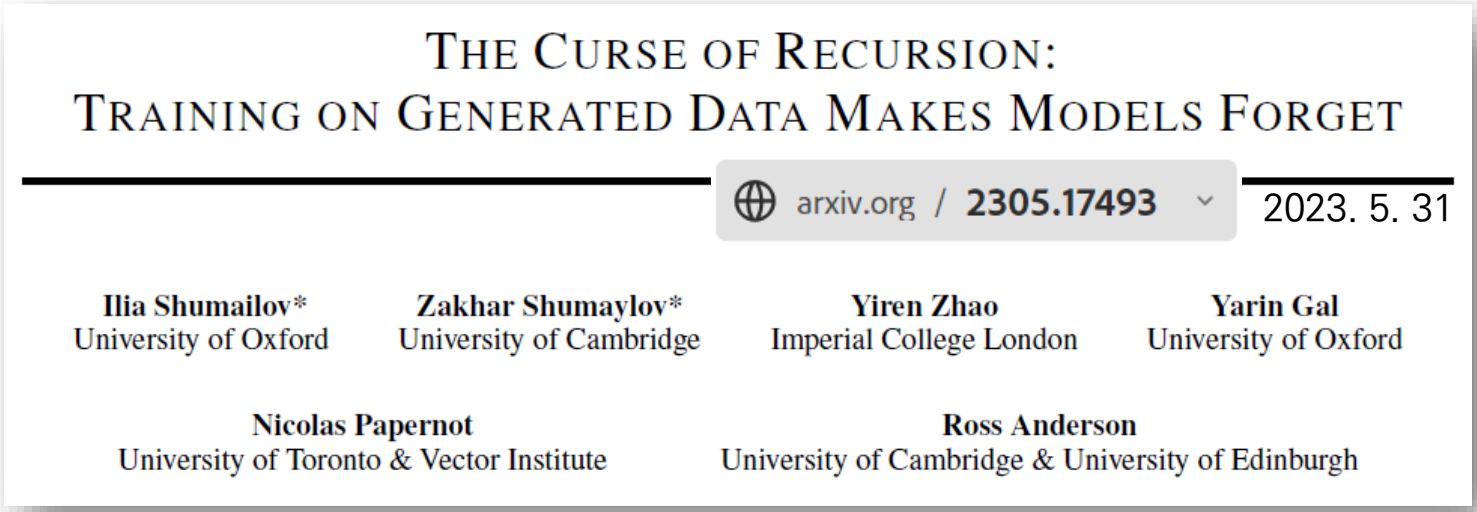
- There is a significant potential for copyright infringement in AI training data, as evidenced by instances of both direct plagiarism and paraphrased content in AI-generated outputs.

European Union's AI Act

- Requirements for General Purpose AI(GPAI)
  - Chapter 5. General-Purpose AI Model ▶ Section 2. Obligations for providers ▶ Article 53

1. Providers of general-purpose AI models shall:
(a) … (b) …
(c) put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;
(d) draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.

Challenge on Synthetic Data



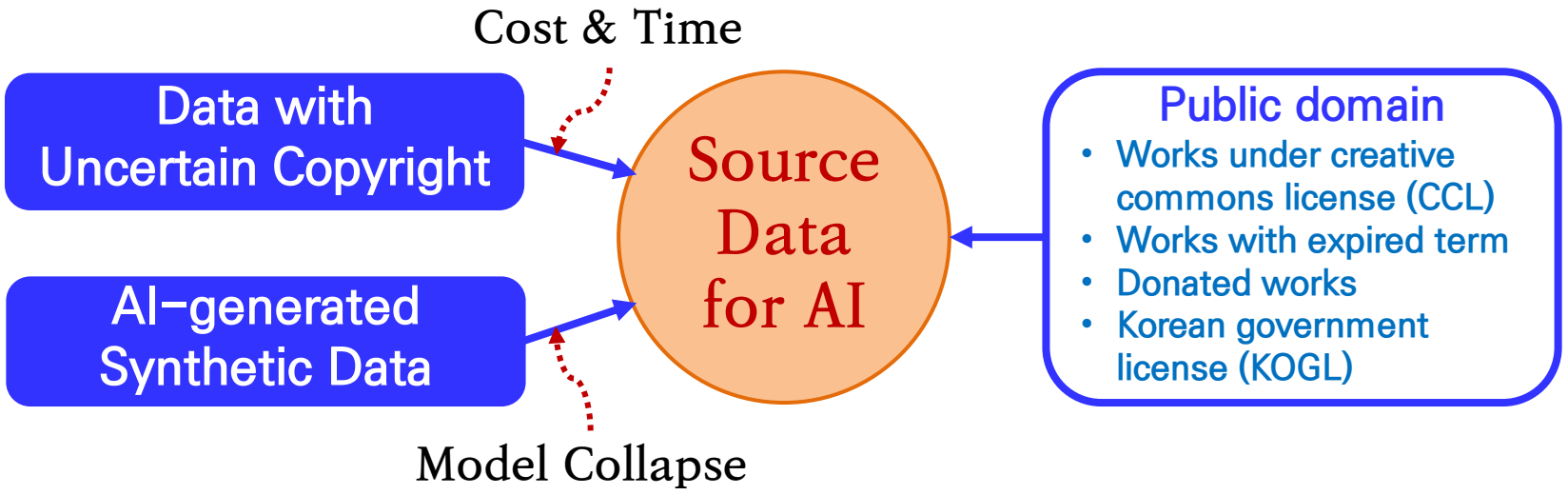
THE CURSE OF RECURSION:
TRAINING ON GENERATED DATA MAKES MODELS FORGET

arxiv.org / 2305.17493 2023. 5. 31

Ilia Shumailov* University of Oxford
Zakhar Shumaylov* University of Cambridge
Yiren Zhao Imperial College London
Yarin Gal University of Oxford
Nicolas Papernot University of Toronto & Vector Institute
Ross Anderson University of Cambridge & University of Edinburgh

- We demonstrate the existence of a degenerative process in learning and name it model collapse;
- We demonstrate that model collapse exists in a variety of different model types and datasets;
- We show that, to avoid model collapse, access to genuine human-generated content is essential.
- The cause of curse: Recursive Learning of AI-generated synthetic outputs

Public domain: New horizons and possibilities as source data for AI



Gong-u-madang
– Using public domain works

89. Variations on the Canon -Gukak

Photos Illustrations Fonts Soundsource Videos Education

Artificial intelligence

Andong Katuri (font)

반듯한 저작권
반듯한 대한민국

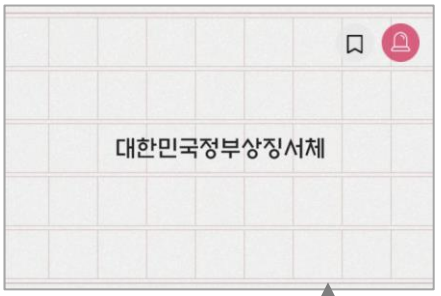
IMAGE

Image AI source data

https://gongu.copyright.or.kr/

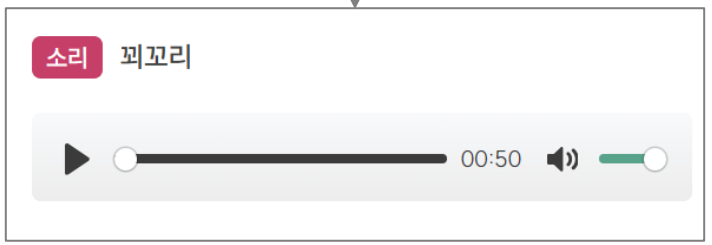
Gonggong-nuri
– Using works owned by public sector

Gonggong-nuri services a total of **28,145,747** works.

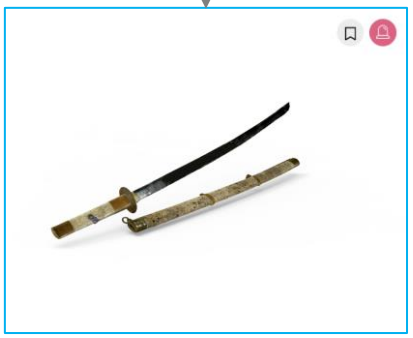


OPEN
공공누리 공공저작물 자유이용허락

Images | AV | Audio | Fonts | 3D | Literary



https://www.kogl.or.kr/



Conclusion

- **The Potential for Copyright Infringement in GenAI's Training Data**
  - High Probability of "Copyright Infringement" in terms of Substantial Similarity and Access
    - Lawsuits emerging due to growing dissatisfaction among copyrights holders.
    - New light needs to be shed on copyright owned by humans
  - Controversy around feasibility of the fair use defense

- **Considerations regarding AI's Training Data**
  - Data with uncertain copyright requires immense time and effort
  - AI-generated synthetic data may cause GenAI model to collapse

- **New horizons and possibilities**
  - Expand works in the public domain and facilitate their use