Volume 7, No. 5 | September–October 2024

RAIL

The Journal of Robotics, Artificial Intelligence & Law

fastcase FULL COURT PRESS

RAIL
The Journal of Robotics, Artificial Intelligence & Law

Volume 7, No. 5 | September–October 2024

THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW (ISSN 2575-5633 (print) /ISSN 2575-5617 (online) at $495.00 annually is published six times per year by Full Court Press, a Fastcase, Inc., imprint. Copyright 2024 Fastcase, Inc. No part of this journal may be reproduced in any form—by microfilm, xerography, or otherwise—or incorporated into any information retrieval system without the written permission of the copyright owner. For customer support, please contact Fastcase, Inc., 729 15th Street, NW, Suite 500, Washington, D.C. 20005, 202.999.4777 (phone), or email customer service at support@fastcase.com.

Publishing Staff
Publisher: Leanne Battle
Production Editor: Sharon D. Ray
Cover Art Design: Juan Bustamante

Cite this publication as:

The Journal of Robotics, Artificial Intelligence & Law (Fastcase)

This publication is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If legal advice or other expert assistance is required, the services of a competent professional should be sought.

A Full Court Press, Fastcase, Inc., Publication

Editorial Office

729 15th Street, NW, Suite 500, Washington, D.C. 20005
https://www.fastcase.com/

POSTMASTER: Send address changes to THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW, 729 15th Street, NW, Suite 500, Washington, D.C. 20005.

Articles and Submissions

Direct editorial inquiries and send material for publication to:

Steven A. Meyerowitz, Editor-in-Chief, Meyerowitz Communications Inc.,
26910 Grand Central Parkway, #18R, Floral Park, NY 11005, smeyerowitz@meyerowitzcommunications.com, 631.291.5541.

Material for publication is welcomed—articles, decisions, or other items of interest to attorneys and law firms, in-house counsel, corporate compliance officers, government agencies and their counsel, senior business executives, scientists, engineers, and anyone interested in the law governing artificial intelligence and robotics. This publication is designed to be accurate and authoritative, but neither the publisher nor the authors are rendering legal, accounting, or other professional services in this publication. If legal or other expert advice is desired, retain the services of an appropriate professional. The articles and columns reflect only the present considerations and views of the authors and do not necessarily reflect those of the firms or organizations with which they are affiliated, any of the former or present clients of the authors or their firms or organizations, or the editors or publisher.

QUESTIONS ABOUT THIS PUBLICATION?

For questions about the Editorial Content appearing in these volumes or reprint permission, please contact:

Leanne Battle, Publisher, Full Court Press at leanne.battle@vlex.com or at 202.999.4777

For questions or Sales and Customer Service:

Customer Service
Available 8 a.m.–8 p.m. Eastern Time
866.773.2782 (phone)
support@fastcase.com (email)

Sales
202.999.4777 (phone)
sales@fastcase.com (email)

ISSN 2575-5633 (print)
ISSN 2575-5617 (online)

Training Artificial Intelligence Models on Synthetic Data: No Silver Bullet for Intellectual Property Infringement Risk in the Context of Training AI Systems—Part 1

Gareth Kristensen, Angela L. Dunning, Gaia Shen,
Prudence Buckland, Jan-Frederik Keustermans, and Alix Anciaux*

This is the introductory part of a multipart series on using synthetic data to train artificial intelligence (AI) models. Part 2 of this series will cover the question of how training AI models on synthetic data may mitigate copyright infringement risks. Part 3 will cover the interplay between synthetic data training sets, the EU Copyright Directive, and the EU AI Act. Part 4 will explore other key legal topics to be considered when using synthetic data to train an AI model.

The recent rapid advancements of artificial intelligence (AI) have revolutionized creation and learning patterns. Generative AI (GenAI) systems have unveiled unprecedented capabilities, pushing the boundaries of what we thought possible. Yet, beneath the surface of the transformative potential of AI lies a complex legal web of intellectual property (IP) risks, particularly concerning the use of "real-world" training data, which may lead to alleged infringement of third-party IP rights if AI training data is not appropriately sourced.

This is because training of GenAI models requires processing of large amounts of data that potentially contain copyrighted works, as well as materials displaying trademarks and data compilations that may be protected by sui generis database rights in the European Union, or other information the use of which may be restricted by contract or terms of use. Only through that training can the AI model be leveraged and applied to generate plausible and human-like new content (such as text, code, images, sound, or video). If

Robotics, Artificial Intelligence & Law / September–October 2024, Vol. 7, No. 5, pp. 365–368.

ISSN 2575-5633 (print) / ISSN 2575-5617 (online).

not adequately deduplicated, filtered, and calibrated, there is also a risk that GenAI systems may generate infringing outputs that are substantially similar to or otherwise replicate (in whole or meaningful part) third-party works protected by copyright.

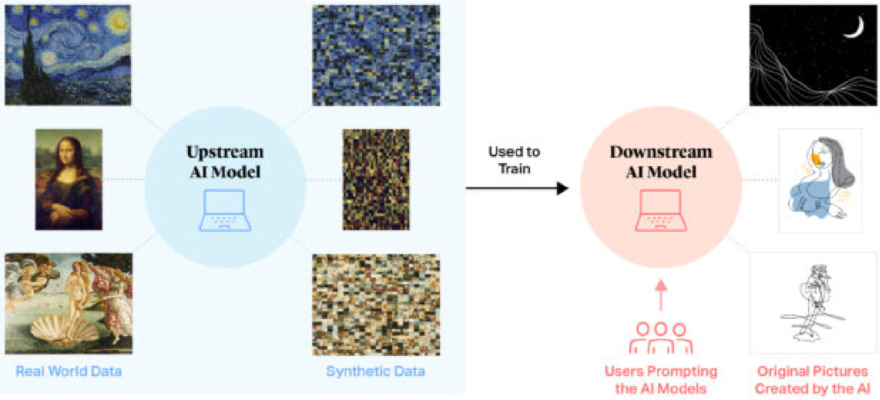
This has given rise to the international debate surrounding how to balance the respective rights and interests of IP rightsholders and AI developers. Several lawsuits have even been launched by rightsholders and representative organizations against developers of GenAI tools, typically claiming that the process of training the AI models utilized by such tools and, in some cases, the output generated by such tools, infringe their IP rights.[1]

Synthetic Data

In this context, "synthetic data" has emerged as a potential solution, as can be seen in Figure 1. Synthetic data comprises data that is artificially generated by an AI model rather than mined or collected from real-world sources and, therefore, should not (in theory) give rise to the same IP infringement risks as using real-world data. Synthetic data mimics real-world data and, if properly developed, should be technically and statistically indistinguishable from such data for the purpose of training AI models.

Several major AI companies are currently using synthetic data to train their AI models.[2] A new type of business has even emerged: companies are now specializing in providing synthetic data sets, either from a pre-existing proprietary database or by creating "bespoke" synthetic data generated on demand for specific

Figure 1.



customers.[3] Synthetic data has many practical use cases already, including in the insurance sector, medical research,[4] or drug discovery and testing.[5]

Synthetic data creates technological, economic, and ethical opportunities, including the potential to:

1. Improve accuracy by mitigating the unreliability of human-made data, which is typically gathered by scraping the erratic web that is the internet;[6]
2. Mitigate or even remove biases and imbalances in existing, human-made data;[7] and
3. 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.[8] In addition, companies are starting to run out of easily accessible, reliable and high-quality real-world data sources to continue training more advanced AI models, thereby increasing demand for synthetic data.[9]

Against this backdrop, we will consider in three future parts of this article whether synthetic data could adequately mitigate IP infringement risks that arise in the context of training AI models under existing and proposed European legal frameworks, with a focus on copyright protection (which has, thus far, emerged as the predominant basis upon which to challenge AI developers).

Notes

[*] The authors, attorneys with Cleary Gottlieb Steen & Hamilton LLP, may be contacted at gkristensen@cgsh.com, adunning@cgsh.com, gshen@cgsh.com, pbuckland@cgsh.com, jkeustermans@cgsh.com, and aanciaux@cgsh.com, respectively.

[1] Some of the most prominent cases currently pending in U.S. courts include: Doe I v. Github, Inc., No. 4:22-cv-06823 (N.D. Cal. Nov. 3, 2022); Andersen et al. v. Stability AI et al., No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023); Getty Images (US), Inc. v. Stability AI, No. 1:23-cv-00135 (D. Del. Feb. 3, 2023). Getty Images launched a similar lawsuit against Stability AI in the United Kingdom. The High Court declined Stability AI's request to dismiss the case in December 2023, and the case will be proceeding to trial in the upcoming months, see https://www.natlawreview.com/article/

getty-images-ai-model-training-lawsuit-uk-against-stability-proceed; J.L. et al. v. Alphabet, No. 3:23-cv-03440 (N.D. Cal. Jul. 11, 2023); Tremblay et al. v. OpenAI, No. 3:23-cv-03223 (N.D. Cal. June 28, 2023); Silverman et al. v. OpenAI, No. 4:23-cv-03416 (N.D. Cal. July 7, 2023); Kadrey et al. v. Meta Platforms, No. 3:23-cv-03417 (N.D. Cal July 7, 2023); Chabon et al. v. OpenAI, No 3:23-cv-04625 (N.D. Cal Sept. 8, 2023); Chabon et al. v. Meta Platforms, No. 3:23-cv-04663 (N.D. Cal. Sept. 12, 2023); Authors Guild v. OpenAI, No. 1:23-cv-08292 (S.D.N.Y. Sept. 19, 2023); Huckabee et al. v. Meta Platforms et al., No. 1:23-cv-09152 (S.D.N.Y. Oct. 17, 2023); Concord Music Group et al. v. Anthropic PBC, No. 3:23-cv-01092 (M.D. Tenn. Oct. 18, 2023); Sancton v. Open AI, Inc. et al., No. 1:23-10211 (S.D.N.Y. Nov. 21, 2023).

[2] These include Microsoft, OpenAI, Cohere, Amazon, and Google (Waymo). See https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de. See also Types of Synthetic Data and 4 Real-Life Examples (2022), https://www.statice.ai/post/types-synthetic-data-examples-real-life-examples#:~:text=Amazon%20is%20using%20synthetic%20data,data%20to%20improve%20fraud%20detection.

[3] For example, see https://scale.com/, https://gretel.ai/, and https://mostly.ai/.

[4] See Mostly AI, "European Commission's JRC: Synthetic Data Will Be the Key Enabler for AI in Europe" (Sept. 15, 2022), reporting on the use of synthetic data to generate cancer data, https://mostly.ai/blog/synthetic-data-for-ai-jrc-report.

[5] On the latter, see EMA reflection paper of July 13, 2023, "The Use of AI in the Medicinal Product Lifecycle," https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle. See also "Synthetic Data Use: Exploring Use Cases to Optimize Data Utility," https://link.springer.com/article/10.1007/s44163-021-00016-y.

[6] MIT, "In Machine Learning, Synthetic Data Can Offer Real Performance Improvements" (Nov. 3, 2022), https://news.mit.edu/2022/synthetic-data-ai-improvements-1103.

[7] Yet, because synthetic data is typically created by an upstream AI model that was fed with real-world data, and synthetic data will in turn be used to train downstream AI models, depending on the exact parameters used to create the synthetic training data set (which may, themselves, contain errors or data bias), synthetic data may also carry risks to transpose or even introduce bias and imbalances.

[8] European Commission, Competition Policy for the Digital Era (2019), p. 73 et seq. See also OECD, Data-Driven Innovation: Big Data for Growth and Well-Being (2015), pp. 391-392.

[9] Financial Times, "Why Computer-Made Data Is Being Used to Train AI Models" (July 19, 2023), https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de.