

Synthetische Daten

Im Zuge der Entwicklungen rund um Big Data und vor allem der weit verbreiteten Nutzung **Künstlicher Intelligenz** (KI) gewinnen Daten immer mehr an Bedeutung und damit an Wert. Zum Training von KI-Algorithmen sind eine Fülle von Daten notwendig. Soweit es sich dabei um personenbezogene Daten handelt, steht dem Gebrauch dieser oft die gesetzliche Beschränkung der Nutzung auf den ursprünglichen Zweck (Zweckbestimmung der DSGVO) entgegen (siehe auch **Digitalisierung und Anonymität**).

Gerade in den Bereichen Gesundheit und Finanzdienstleistungen sind die Originaldaten sehr sensibel, aber auch in anderen Bereichen wird verstärkt auf die Weiterverarbeitung von einmal generierten Daten Wert gelegt. Um dem Problem der Beschränkung nach der DSGVO zu entgehen, wird mittlerweile versucht, mit so genannten „synthetischen Daten“ (Replica Analytics 2020) zu operieren. Diese werden künstlich und mit Hilfe von KI erzeugt, um sie dann wiederum als Trainingsdaten für KI-Systeme zu nutzen. Sie bilden ein Daten-Set, das dieselben statistischen Merkmale aufweisen soll, wie die zugrundeliegenden Originaldaten und so als Proxy für das Training des Algorithmus genutzt werden können.

Dieses Versprechen macht synthetische Daten zu einem wertvollen Instrument, um andere Anonymisierungstechniken zu unterstützen bzw. abzulösen. Allerdings gibt es auch Einwände gegen die Nutzung synthetischer Daten.¹ Zum einen wird ihre Anonymisierungsstärke hinterfragt (Stadler et al. 2022) zum anderen ist die Technik noch nicht ausgereift und sehr voraussetzungsvoll² und bedarf sehr hoher Kompetenz der Durchführenden. Weiters entsteht durch die Überführung von Daten in ein Modell immer eine Komplexitätsreduktion gegenüber der Wirklichkeit. Die Vorstellung, dass ein Modell eines Daten-Sets (ebenfalls bereits nur ein Ausschnitt der Realität) erstellt wird, auf dessen Basis ein Modell der Wirklichkeit erzeugt wird, aufgrund dessen dann Entscheidungen für Menschen in der Realität getroffen werden, lässt erahnen, wie wichtig hier die Kompetenz der Beteiligten, aber auch die ethischen Kriterien für diese sein müssen. Es sollte genau überlegt werden, in welchen Bereichen diese Technologie eingesetzt werden kann.

Grundsätzlich sind synthetische Daten „fake data“ und auch die Basis für sogenannte **Deep Fakes**, z. B. synthetische Bilder von Menschen, die es nicht gibt,³ oder Videos, in denen bekannte Menschen etwas tun oder sagen, was sie nicht getan haben. Eine Entwicklung, die besonders problematisch erscheint und die Frage von Echtheit und Wahrheit im Digitalen neu stellt. Positive Anwendungen von synthetischen Daten könnten etwa darin liegen, dass bestimmte Verzerrungen (biases), die in Originaldaten

¹ diginomica.com/data-managers-beware-synthetic-data-still-has-limitations.

² universityaffairs.ca/news/news-article/the-promise-and-pitfalls-of-synthetic-data/.

³ thispersondoesnotexist.com.

vorhanden sind, ausgeglichen werden können. So wurden etwa synthetische Daten von afrikanischer Kleidung erzeugt, um Computer-Vision-Algorithmen damit zu trainieren. Bisher gab es dazu nur Daten westlicher Kleidung.⁴

Mittlerweile gibt es eine Reihe von Anbietern, die derartige Daten oder auch Prozesse anbieten. Hervorzuheben ist hier, dass ein österreichisches Start-up⁵ international sehr erfolgreich positioniert zu sein scheint. Ergänzend zum Konzept der synthetischen Daten gibt es zum Schutz der Privatsphäre auch den Ansatz der „differential privacy“, bei der nichts erfunden wird, sondern die Originaldaten in Datenbanken so verändert werden, dass Angreifer bei missbräuchlicher Verwendung nicht sicher sein können, dass eine bestimmte Person in der Datenbank enthalten ist. Dazu forscht in Österreich Monika Henzinger.⁶

Zitierte Literatur

Replica Analytics, 2020, Synthesis Tutorials;

replica-analytics.com/synthesis-tutorials.

Stadler, T., Oprisanu, B. und Troncoso, C., 2022, Synthetic Data – Anonymisation Groundhog Day, *arXiv:2011.07018 [cs.LG]*.

⁴ technologyreview.com/2022/02/23/1044965/ai-synthetic-data-2/.

⁵ mostly.ai.

⁶ informatik.univie.ac.at/news-events/beitrag/news/2-erc-grant-fuer-prof-dr-monika-henzinger/.