AI & EQUALITY

Synthetic data:

Potential and Challenges for AI & Equality

By Sofia Kypraiou

AI & Equality Toolbox Data Science Lead / Women At The Table

October 2023

I.What is Synthetic Data

Modern machine learning algorithms, especially deep learning models, require huge amounts of training data to accurately learn and generalize patterns from data. Real-world examples, while valuable, are often not enough or available in quantity, as these algorithms need to be exposed to a wider range of data.

This is where synthetic data, a concept dating back to the 1970s but currently experiencing a surge in popularity due to advancements in AI and modeling, can become a valuable tool.

What exactly is synthetic data?

Synthetic data is artificially generated data designed to mimic real-world characteristics without containing any actual information. This versatile resource comes in various forms, including text, media (video, images, sound), and tabular data, and it's shaping the future of data science.

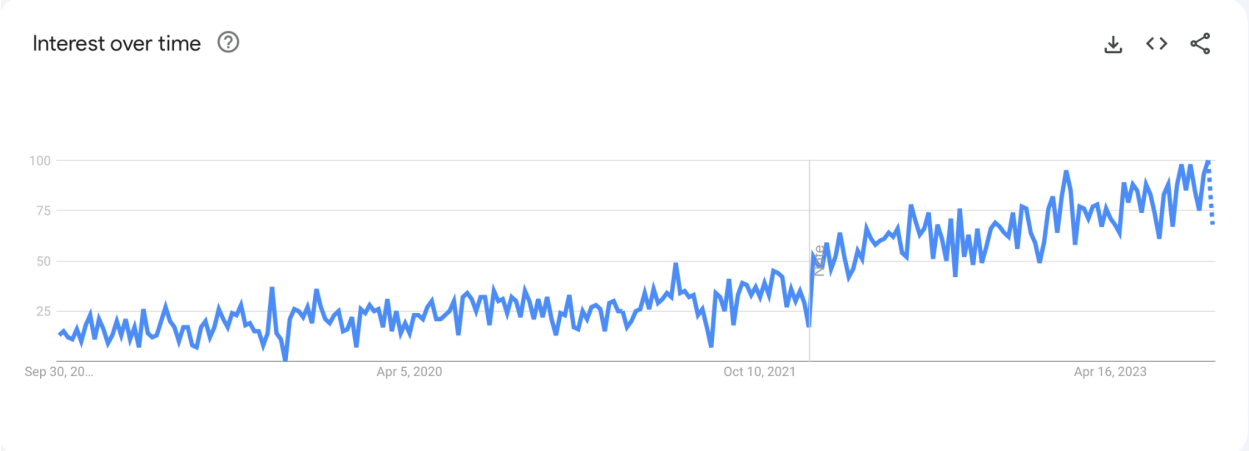


Image source: World-wide Google trends on the term 'synthetic data' over the last 5 years (since September 2018), where we observe its increasing popularity.

By 2024, Gartner predicts 60% of data for AI will be synthetic to simulate reality, in comparison to 1% in 2021, showing how essential the role of synthetic data will be

in future applications[1]. But before jumping into generating more and more data, it's important to take a step back and reflect - otherwise we run the risk of continuing to create biased data. What specific role do the new synthetic data play, and how do they contribute to the broader goals of fairness and equity?

The incorporation of synthetic data should be a carefully considered step, as its ethical and responsible application is vital. Governance and vigilance about biases are essential to prevent this data from suffering the same challenges as organic data.

[1] 'Three Factors Weighing on Growth Rates in 2023'. Gartner, https://www.gartner.com/en/insights. Accessed 3 Oct. 2023.

II. Benefits of Synthetic Data: Privacy, Customization, and Efficiency

One of the most significant advantages of synthetic data is privacy. By generating data without revealing personal information, it enables safe analysis of sensitive domains such as healthcare and finance. This allows researchers to work with data that maintains the statistical properties of real patient data without compromising patient privacy, and can avoid cyber and black-box attacks where models infer the details of training data. Additionally, the ability to create data on-demand, customized to specific requirements, and in large quantities offers immense benefits.

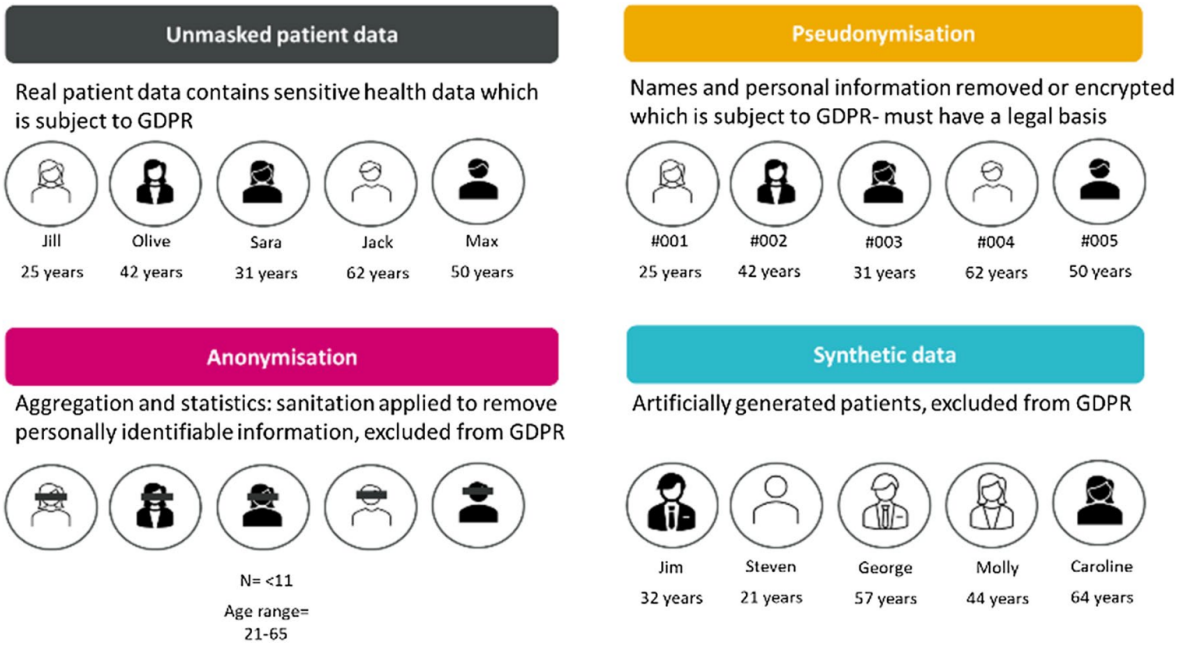


Image source: James, Stefanie, et al. 'Synthetic Data Use: Exploring Use Cases to Optimise Data Utility'. Discover Artificial Intelligence, vol. 1, no. 1, Dec. 2021, p. 15. DOI.org (Crossref), https://doi.org/10.1007/s44163-021-00016-y.

Synthetic data finds extensive application in testing, improving model performance, and reducing expense. It proves cost-effective for testing applications and can significantly boost model accuracy by expanding datasets through various transformations (for example style transfer techniques in generating synthetic images) and reducing the time and resources needed for data collection and cleaning.

Moreover, synthetic data can be employed for model training in scenarios where acquiring real-world data is challenging or expensive, saving organizations time and resources. In the healthcare scenario, synthetic data can be used to generate realistic images of organs or tissues, which can then be used to train algorithms to recognize patterns and detect abnormalities in real patient images. It can enable more accurate and personalised medical care, enhancing clinical information to patients.[2]

Additionally, synthetic data can augment incomplete or missing datasets to enrich analytics and train machine learning models, improving the fairness[3] in the source of the data. When created with the principles of inclusion, they provide an opportunity to balance underrepresented groups by creating additional data points.[4] The ability to create data with known characteristics is particularly advantageous for testing AI applications, fine-tuning algorithms, and developing robust machine learning models.

---

[2] HealthManagement.org, et al. 'Radiology Management, ICU Management, Healthcare IT, Cardiology Management, Executive Management'. HealthManagement. https://healthmanagement.org/c/healthmanagement/issuearticle/how-imaging-generative-ai-will-transform-the-medical-radiological-practice . Accessed 12 Oct. 2023.

[3] Navarro, Madeline, et al. Data Augmentation via Subgroup Mixup for Improving Fairness. arXiv, 13 Sept. 2023. arXiv.org, https://doi.org/10.48550/arXiv.2309.07110 .

[4] Brown, Annie. 'Synthetic Data Promises Fair AI And Privacy Compliance, But How Exactly Does It Work?' Forbes, https://www.forbes.com/sites/anniebrown/2020/12/17/synthetic-data-promises-fair-ai-and-privacy-compliance-but-how-exactly-does-it-work/ . Accessed 2 Oct. 2023.

III. Pitfalls of Synthetic Data: Biases and Challenges

However, the improper generation of synthetic data can amplify bias. If the algorithms fail to accurately capture the distribution of real-world data, the synthetic data may inherit biases, leading to skewed models and predictions. Moreover, synthetic data may not adequately cover outliers or rare events, which are crucial in certain applications.

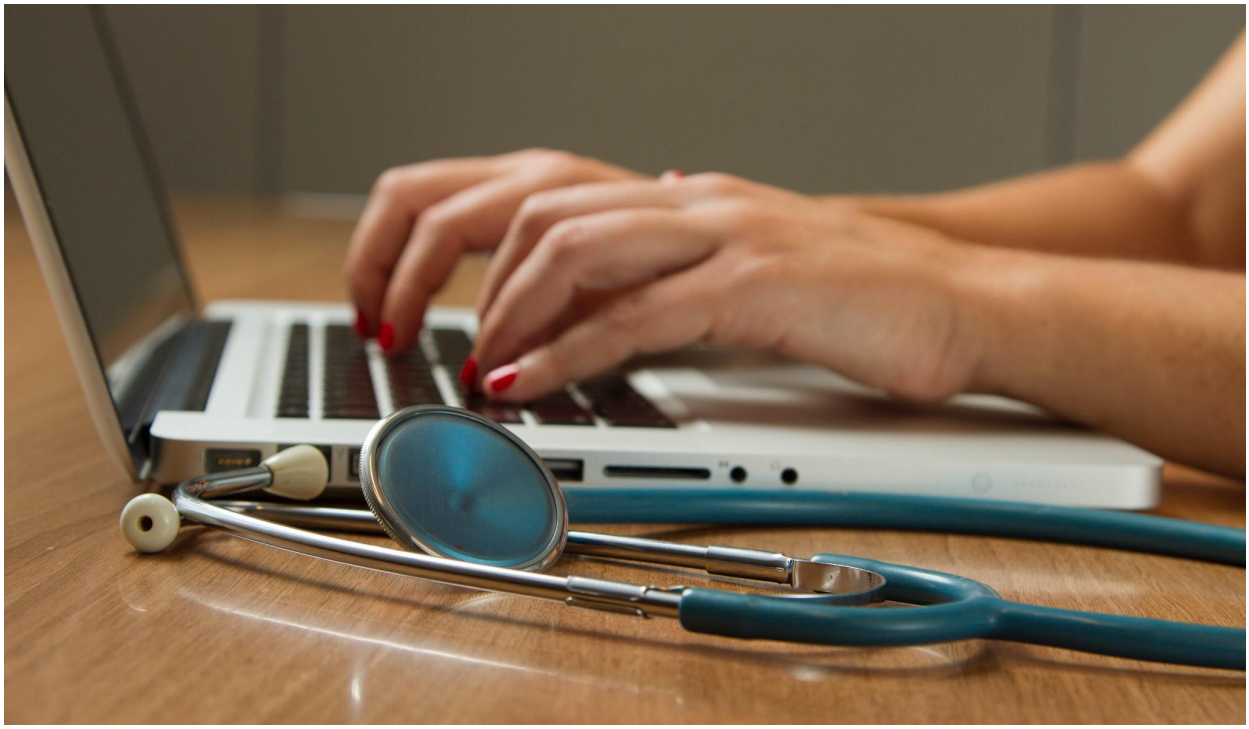


Image source: Photo by National Cancer Institute on Unsplash

In healthcare, some medical conditions or diseases are extremely rare, but they can have significant consequences if undetected. If a diagnostic algorithm is trained on synthetic data that doesn't accurately represent the occurrence of these rare diseases, it could lead to missed diagnoses.

Image source: Photo by ThisIsEngineering: https://www.pexels.com/photo/code-projected-over-woman-3861969/

Synthetic data can only mimic real-world data; it is not a replica. Therefore, synthetic data may not cover some outliers that original data has.

Creating high-quality synthetic data is complex, requiring advanced machine learning knowledge and significant computational resources. Training AI models on raw, unfiltered synthetic data can lead to detrimental consequences, including "irreversible defects"[5], as researchers from universities including Oxford and Cambridge recently warned. Based on their research, training AI models using their own unfiltered model-generated content that might include inaccuracies, false information, or even fabricated data, leads to a degenerative process whereby, over time, models forget the true underlying data distribution; meaning that we start losing information about the original data. In simpler terms, relying on unverified AI-generated data could harm the long-term effectiveness of those AI systems.

[5] Shumailov, Ilia, et al. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv, 31 May 2023. arXiv.org, https://doi.org/10.48550/arXiv.2305.17493 .

IV. Synthetic Data in Healthcare: Opportunities, Diversity, and Ethical Considerations

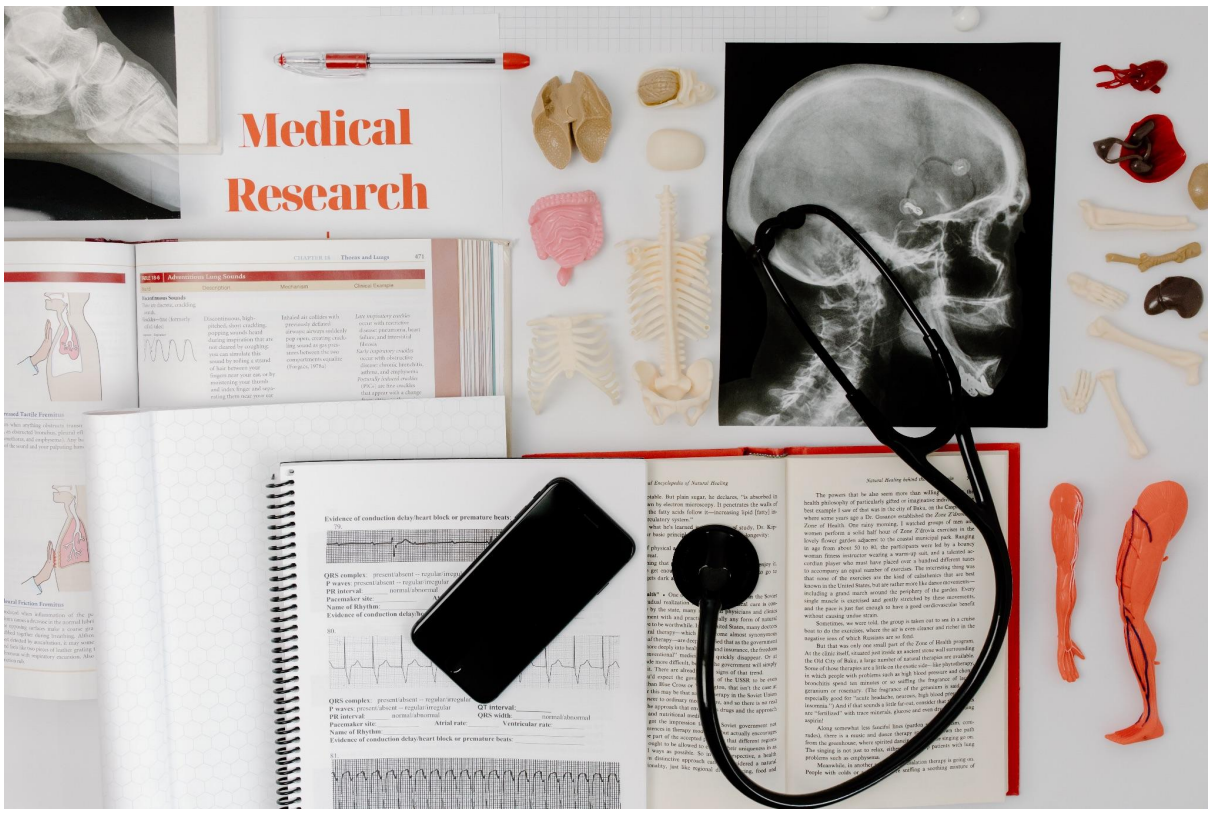


Image source: Photo by Tara Winstead from Pexels: https://www.pexels.com/photo/black-usb-cable-on-white-and-red-box-7723388/

In healthcare, synthetic data presents both opportunities and challenges. Before using synthetic data, many patient cohorts had minimal participation. Statistics show that racial and ethnic minorities comprise 39% of the United States population but only account for 2% to 16% of clinical trial participants[6] Factors like age, biological sex, disabilities, chronic comorbidities, geographical location, gender identity, race, and ethnic background, may influence how an individual reacts to a certain drug, medical device, or treatment plan. If patients in clinical trials do not

[6] DeArment, Alaric. 'As Precision Medicine Grows, so Does the Importance of Clinical Trial Diversity'. MedCity News, 7 July 2019, https://medcitynews.com/2019/07/as-precision-medicine-grows-so-does-the-importance-of-clinical-trial-diversity/.

represent the whole community, there is the risk that differences in drug metabolism, side effect profiles, and outcomes will be missed.

This also translates when using synthetic data. The lack of diversity in synthetic patient cohorts can result in AI models that perform poorly on real-world populations.

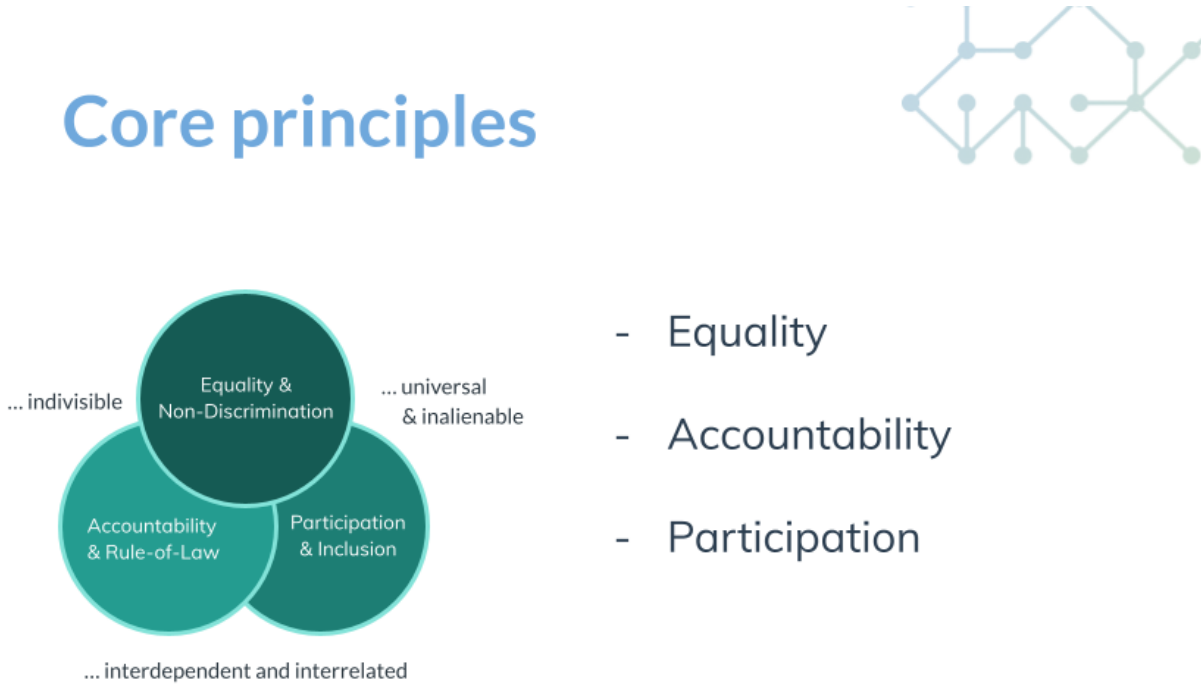
As an example, generating data for 500 Black male patients and 500 Black female patients using a synthetic data generator trained on predominantly white medical records would not accurately reflect the true disease progression and outcomes experienced by Black patients.[7]

To address this, representative real-world data must be collected first to ensure that AI models do not perpetuate healthcare disparities.

Moreover, the synthetic data landscape in healthcare is fraught with ethical considerations. While synthetic data offers the potential to accelerate medical research, drug development, and personalized treatment strategies, it must be used with care to avoid reinforcing biases and ensuring patient privacy and consent.

[7] Talby, David. 'Council Post: The Dangers Of Using Synthetic Patient Data To Build Healthcare AI Models'. Forbes, https://www.forbes.com/sites/forbestechcouncil/2023/05/26/the-dangers-of-using-synthetic-patient-data-to-build-healthcare-ai-models/. Accessed 16 Oct. 2023.

V. Strategies for A Human Rights-based Use of Synthetic Data



To ensure the responsible and ethical use of synthetic data, we must adopt a Human Rights-Based Approach, guided by core Human Rights principles of inclusion, participation, and non-discrimination. This approach addresses the challenges posed by synthetic data and its policy implications.

In the context of ensuring fairness in synthetic data, it's important to initiate the process from the very beginning of your objectives. Just because the capability exists to use synthetic data doesn't automatically mean that you should. It's crucial to carefully consider the motives behind employing or generating synthetic data. What specific use case does it serve, and how will it contribute to fairness and equity? Is the incorporation or creation of synthetic data an essential step in achieving fairness objectives, and if so, why?

VI. Policy Considerations and Evolving Guidelines

As technology firms introduce generative AI applications, policymakers worldwide are wrestling with the associated challenges, while researchers are debating the management of generative AI, spanning from mitigation strategies during model design and development to its market introduction and beyond. The public discourse surrounding synthetic data, and by extension the generative AI techniques that are commonly used to create synthetic data are just emerging, as the Organisation for Economic Co-operation and Development (OECD) reports in recent "Initial policy considerations for generative artificial intelligence"[8].

In these steps, recommendations have been made to employ synthetic data in policy-making processes. The Administrative Data Research UK[9] proposes using lo-fi synthetic data across government and researchers to reveal whether data for a given policy is available and usable; for writing and testing code before access to real-world data is available; and to provide quicker access where there are data security issues.

The UK Statistical Office has also released a series of ethical principles and a related self-assessment tool[10] for scientists to prioritize the public good, data confidentiality, risk awareness in new methods, legal compliance, public acceptability, and data transparency.

[8] Lorenz, P., K. Perset and J. Berryhill (2023), "Initial policy considerations for generative artificial intelligence", OECD Artificial Intelligence Papers, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en.
[9] Calcraft, P., I. Thomas, M. Maglicic and A. Sutherland (2021) 'Accelerating public policy research with synthetic data', 14 December, Behavioural Insights Team, https://www.adruk.org/fileadmin/uploads/adruk/Documents/Accelerating_public_policy_research_with_synthetic_data_December_2021.pdf
[10] 'Ethics Self-Assessment Tool'. UK Statistics Authority, https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/. Accessed 16 Oct. 2023.

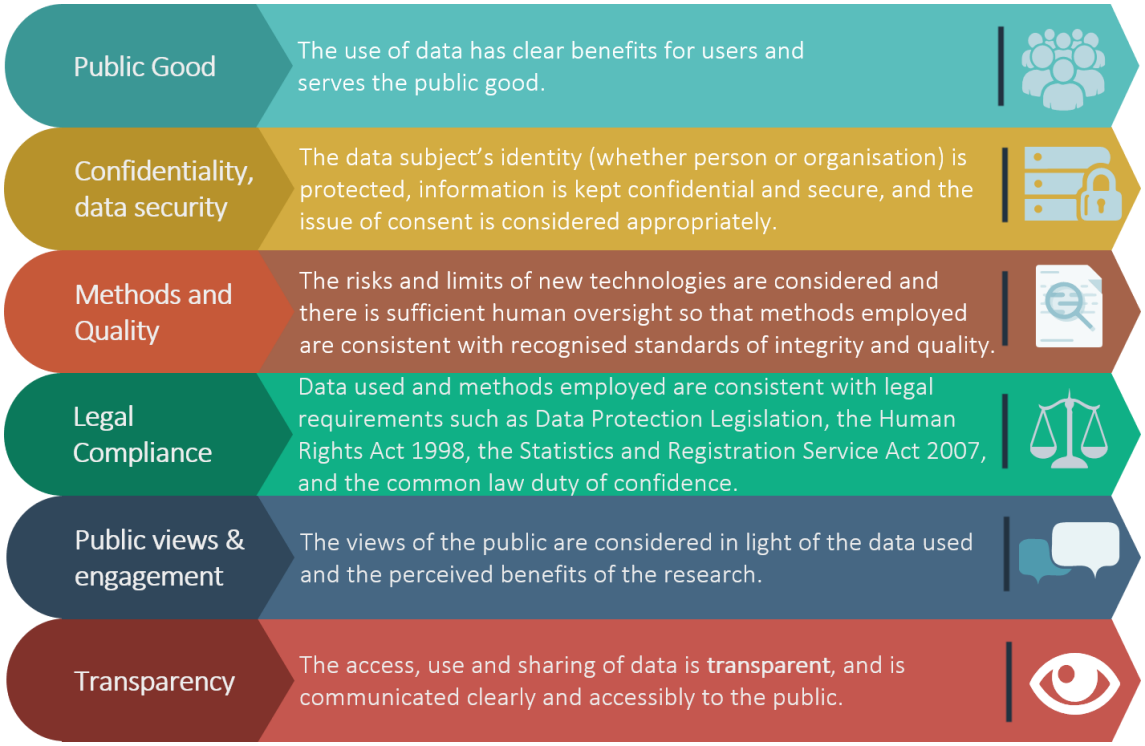


Image source: 'Ethical Principles'. UK Statistics Authority, https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethical-principles/. Accessed 9 Oct. 2023.

Synthetic data doesn't need just technical capability but also knowledge and understanding of privacy requirements and risks. Embedding documentation about utility, transparency, and an audit trail is essential to establish trust.[11]

Organizations like the Partnership on AI (PAI)[12] provide recommendations for responsible practices in synthetic media development and deployment. Their practices are the result of feedback from more than 100 global stakeholders, with representatives from industry, civil society, media/ journalism, and academia. detailed practices for each stakeholder category to promote ethical and responsible behavior in the development, creation, and distribution of synthetic media.

[11] James, S., C. Harbron, J. Branson and M. Sundler (2021) 'Synthetic data use: exploring use cases to optimize data utility', Discover Artificial Intelligence, 1 (15)

[12]

From a more technical perspective, applying fairness constraints to synthetic data, similar to traditional machine learning datasets, can enhance diversity and balance in synthetic data.

MIT researchers release the Synthetic Data Vault, a set of open-source tools meant to expand data access without compromising privacy. IBM[13] researchers have developed a platform for monitoring the trustworthiness of synthetic data, connecting various stakeholders and [14]promoting transparent reporting.

[13] Belgodere, Brian, et al. Auditing and Generating Synthetic Data with Controllable Trust Trade-Offs. arXiv, 2 May 2023. arXiv.org, https://doi.org/10.48550/arXiv.2304.10819.
[14] The real promise of synthetic data | MIT News | Massachusetts Institute of Technology

VII. The Path Forward: Toward a Responsible and Equitable Future

In conclusion, synthetic data is transforming data science by addressing privacy concerns, enabling customization, and reducing costs, in a very rapid way.

However, it comes with its own set of challenges, particularly in terms of bias. By embracing ethical practices, transparency, and fairness constraints, we can harness the potential of synthetic data and create a more responsible and equitable dataset for the future of data science.

About the author

Sofia Kypraiou, a data scientist with a specialization in ethics and human rights, earned her MSc in Data Science from the École Polytechnique Fédérale de Lausanne (EPFL). She holds a BSc in Computer Science from the National and Kapodistrian University of Athens.

She developed the technical components of the workshop, "<AI & Equality>: A Human Rights Toolbox", during her MSc thesis at EPFL and works with Women At The Table, in collaboration with the Office of the United Nations High Commissioner for Human Rights (OHCHR). This workshop merges the domains of data science and human rights through a critical analysis methodology.

She has delivered the <AI & Equality>: A Human Rights Toolbox workshop at various universities and art festivals, contributing to the ongoing dialogue at the intersection of these domains.

Resources

- 'Ethical Principles'. UK Statistics Authority, https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethical-principles/ . Accessed 9 Oct. 2023.
- Shumailov, Ilia, et al. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv, 31 May 2023. arXiv.org, https://doi.org/10.48550/arXiv.2305.17493.
- Talby, David. 'Council Post: The Dangers Of Using Synthetic Patient Data To Build Healthcare AI Models'. Forbes, https://www.forbes.com/sites/forbestechcouncil/2023/05/26/the-dangers-of-using-synthetic-patient-data-to-build-healthcare-ai-models/ . Accessed 11 Oct. 2023.
- Lorenz, P., K. Perset and J. Berryhill (2023), "Initial policy considerations for generative artificial intelligence", OECD Artificial Intelligence Papers, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en.
- Belgodere, Brian, et al. Auditing and Generating Synthetic Data with Controllable Trust Trade-Offs. arXiv, 2 May 2023. arXiv.org, https://doi.org/10.48550/arXiv.2304.10819.
- 'PAI's Responsible Practices for Synthetic Media'. Partnership on AI - Synthetic Media, https://syntheticmedia.partnershiponai.org/ . Accessed 11 Oct. 2023.
- Government of Canada, Statistics Canada. Unlocking the Power of Data Synthesis with the Starter Guide on Synthetic Data for Official Statistics. 1 Mar. 2023, https://www.statcan.gc.ca/en/data-science/network/synthetic-data.

Tools:

- The Synthetic Data Vault. Put Synthetic Data to Work! https://sdv.dev/ . Accessed 11 Oct. 2023.