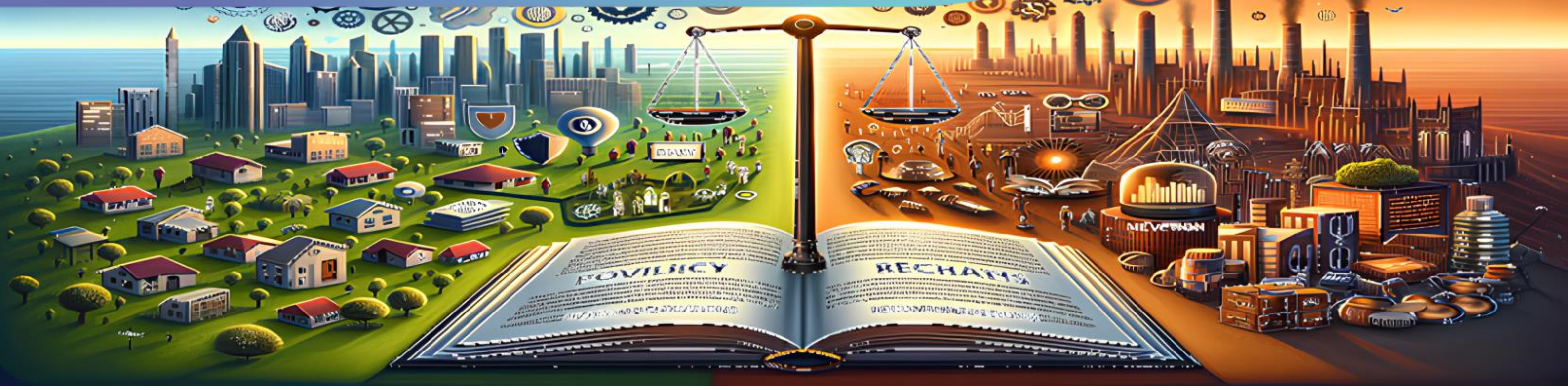
European Commission

SCIENCE FOR POLICY BRIEF



AI Generated Synthetic Data in Policy Applications

This policy brief delves into the field of AI-generated synthetic data and its transformative potential in policy-making. It explores the innovative synthesis methods that create data replicas, which protect privacy and facilitate research by maintaining statistical features of real-world phenomena without exposing sensitive information. The document outlines the shift from traditional data handling to advanced generative AI techniques, highlighting the benefits and challenges associated with this transition. It emphasizes the role of synthetic data in breaking down silos and enhancing privacy, while also discussing the risks of biases and the need for robust quality controls. The brief introduces agentic AI models which leverage large language models to simulate human-like behaviours in policy scenarios. The document underscores the necessity for expert knowledge, collaborative research, and substantial investments to harness the full potential of generative AI and synthetic data in informed decision-making and policy-making. The goal is to provide to policymakers, researchers, and other stakeholders, valuable insights into the future of data-driven policy support and the role of synthetic data in shaping effective and ethical governance.

HIGHLIGHTS

→ **Synthesis methods** use either pre-trained / fine-tuned sources for image or text generation, or training on structured, often tabular, datasets to create replicas of real-world phenomena.

→ Synthetic replicas protect privacy by replicating data's statistical features, avoiding the exposure of sensitive real-world data. They open up silos, remove privacy burden from sensitive data and bring knowledge into domains such as AI and healthcare to **facilitate research and product development**.

→ **Large language and multimodal models** contain pre-trained world models that are capable of generalization and exploratory guidance, and foster creative processes.

Synthetic data risks include **quality issues and biases**, requiring robust quality controls and strategies to prevent skewed outcomes. Their use raises security and ethical issues, requiring strict measures and guidelines to prevent misuse and ensure data integrity.

Joint Research Centre

REAL, SYNTHETIC, GENERATIVE: BRIDGING DATA GAPS AND EHANCING PRIVACY

Data evolution: from isolation to integration

How can we unlock the potential of data that is currently isolated, and how can we use this data while still protecting individual privacy? Generative Artificial Intelligence (AI) provides innovative solutions to these longstanding challenges.

Real individual data is a valuable asset for understanding the world around us. However, it is often underused, locked into databases, and can lose its authenticity when it is combined with other data or processed to protect privacy. Combining data from different sources can be complex and requires a deep understanding of the data itself.

In recent years, new tools have been developed to overcome these obstacles. Data is essentially a snapshot of reality, and the way we collect and interpret it can sometimes distort the true picture. Synthetic data, which are artificially created replicas of real data, introduce[1][2] controlled distortions that we can measure. This allows us to avoid the loss of detail that happens when we combine real data sets, and it helped us keep the important patterns and trends intact. Synthetic data is simpler than data produced by Generative AI, which is more complex and requires careful prompting to extract useful information and to identify any inaccuracies, often referred to as “hallucinations”.

Using generative AI can be quick and convenient, but also comes with its own set of challenges. Currently, there are no established rules or ethical guidelines for using this kind of data in policy making, which means that adopters need to be well informed about its potential risks.

From precision to big picture analysis

The advent of Big Data era marked a shift from isolated data sets to a large, interconnected network of data that is updated almost instantly. Big data’s sheer volume allows for quick correction of errors or unusual data points. Similarly, AI systems, such as neural networks, learn from a wide range of data and can provide insights when they make predictions. Moving from precise, exact data to a more probabilistic approach may reduce the level of detail, but also gives us a more comprehensive and nuanced view of the world.

Precise data sets often require extensive standards to ensure they can work together, but probabilistic data is easier to combine with other types of data. Generative AI models, which are trained on large amounts of data, are good at making general predictions. This makes it easier to integrate them with other probabilistic data sets, such as those that track human behaviour, weather patterns, traffic, or natural events.

QUICK GUIDE

Three approaches for data-driven policy support

1. **Traditional data** involve conventional, sometimes siloed datasets offering specialised yet non-comprehensive snapshots of reality. This data is transparent and traceable, ensuring that properties are verifiable and grounded.
2. **Synthetic replicas** provide analysts with detailed insights from data while maintaining privacy, breaking down silo barriers without compromising privacy. The replication is created algorithmically, and the data are intended to be used as a stand-in for operational or production data. The primary purpose of synthetic data is to protect privacy or to create data sets for testing or training models when real data is scarce or sensitive. Synthetic data should maintain the statistical properties of the original dataset, allowing for the preservation of critical data distributions without exposing the actual data.
3. **Fully AI-generated data** usually refer to data that are produced by artificial intelligence models, today typically by large language models. They compress massive corpora of text and data into their world models and provide targeted information at generation. The sheer amount of amassed knowledge enable the large language models to generate contextualized information from a specific view point and deeper insights into underlying causes, such as behavioural patterns. Yet, these models are black boxes still lacking explainability, making scrutiny challenging.

Agentic AI models: simplifying complexity for policy insights understanding AI agents and their role in policy simulation

Artificial Intelligence has evolved to not only answer complex questions, but also to simulate real-world interactions through AI agents[7]. These agents, working collaboratively, aim to replicate human decision-making processes. Common frameworks like BabyAGI and AgentGPT are designed with the ambitious goal of achieving Artificial General Intelligence (AGI), where AI can perform any intellectual task that a human can.

These frameworks are adept at drawing from diverse data sources, including scientific literature and private databases, to create a synthesised understanding of various topics. By doing so, they provide a rich context for analysis and decision-making.

The power of open-ended AI agents in societal Modelling

Open-ended AI agents, powered by Large Language Models (LLMs), have the potential to revolutionise how we model societies. These agents can be based on real or synthetic populations and are programmed to mimic human behaviour. LLMs, which are trained on extensive collections of text, enable these agents to act in ways that resembles human responses[8][9]. Unlike traditional models, AI agents[3] can adapt to complex changes in their environment[10]. They can process the impact of multiple factors, such as policy changes, natural events, or economic fluctuations, and adjust their behaviour accordingly. This adaptability, represents a significant shift in modelling techniques, offering a more dynamic and integrated approach to understanding complex systems.

EU Policy context: Data Strategy and legal framework

The European Strategy for Data aims to maximise the benefits of data and AI while upholding privacy, security and EU values. It promotes a unified data market, encouraging the sharing of data across borders and sectors, which is essential for integrating data sets and utilising AI in analytics and decision-making.

Key legal instruments like the Data Governance Act (DGA) and the Data Act have been developed to foster trust in data sharing and to ensure fairness in the data economy. These acts regulate data access and usage, aiming to prevent data monopolisation and empower users with more control over their data.

In addition the proposed EU Artificial Intelligence Act introduces a risk-based approach to AI development, ensuring ethical and responsible use. In this context, synthetic data generated by LLMs could serve as a valuable complement or alternative to traditional data sources.

Challenges and opportunities of synthetic data use

The adoption of synthetic data and generative AI is not without challenges. Ensuring the accuracy, quality, and unbiased nature of this data is critical and requires thorough validation processes.

Ethical considerations, particularly around misuse, are the forefront of these concerns. Current legal frameworks have yet to catch up with these advancements, potentially leaving a gap in governance that needs to be addressed.

For organizations to effectively leverage generative AI and synthetic data, they must have access to specialised knowledge. This presents a barrier for those without the necessary technical expertise. To support the computational demands of advanced AI models, significant investments in technology and infrastructure are essential. Initiatives like the EuroHPC[4] are instrumental in providing the necessary resources.

Developing universal standards for the creation, management, interpretation and sharing of synthetic data is vital. These standards will ensure that data is interoperable and reliable across various systems and sectors, facilitating broader and more effective use.

Opportunities for policy support

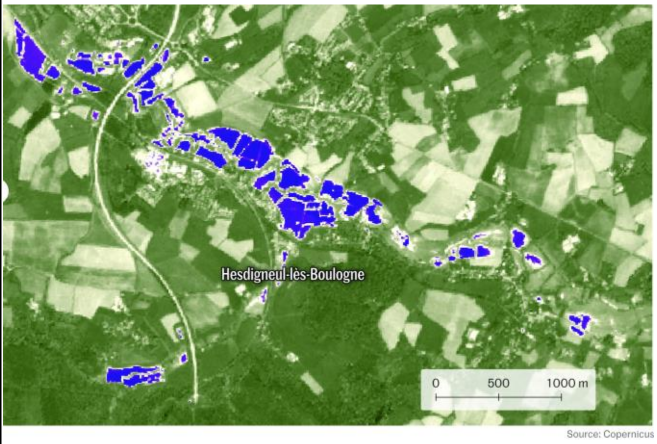
In the EU policy context, generative AI and synthetic data have the potential to enhance the accuracy and depth of policy simulations. This in turn leads to improved forecasting and decision-making. The use of these technologies goes beyond creating realistic imitations, such as deep fakes. By generating synthetic versions of data from multiple sources, we can integrate previously isolated data sets. This integration is crucial for computational models that support policy development, allowing for greater academic evaluation of the outcomes.

Generative AI also presents new research opportunities by enabling the exploration of scenarios that traditional data sources cannot capture. Overcoming the current limitations, synthetic data could democratize data access, empowering smaller organizations to engage more actively in data-driven policy-making.

A Comprehensive Approach to Harnessing AI and Synthetic Data

To fully realise the transformative potential of generative AI and synthetic data, a comprehensive strategy is required. This strategy should include the development of expert knowledge and the establishment of robust frameworks to rigorously evaluate the quality and reliability of data. Such frameworks are necessary to ensure the data's relevance and applicability in policy-making. A collaborative research effort, supported by the Member States of the European Union, is crucial for developing a shared understanding and methodology for integrating these technologies into policy-making. Additionally, significant investments are needed to advance both the technological capabilities and the human expertise required to navigate and innovate in this field. These investments will contribute to informed and effective policy-making, leveraging the full potential of generative AI and synthetic data.

Box 1 – Example of an agent use for disaster response



In a recent case study conducted for the November 6-14, 2023 flooding involving municipalities in Pas-de-Calais[5], AI agents created from synthetic data based on population Census statistics demonstrated potential utility in emergency management simulations. These LLM-driven agents, designed to mirror the demographic and behavioral patterns of real populations, were tested in flood evacuation scenarios. Their responses closely paralleled actual human behaviors recorded during emergencies, particularly when agents were timely informed about worsening weather conditions through integration of historical data from Open Meteo[6]. This simulation has highlighted the potential of AI-driven models in improving the realism and effectiveness of emergency response drills, suggesting that such technologies can enhance the preparedness and adaptive strategies in crisis situations. This case supports the broader application of advanced AI simulations in policy-making and emergency management across the European Union, and demonstrates advantages of integration of observed or modelled data with AI agents.

Source: authors' own elaboration

REFERENCES

[1] Hradec, J., Craglia, M., Di Leo, M., De Nigris, S., Ostlaender, N., & Nicholson, N. (2022). Multipurpose synthetic population for policy applications. *No. JRC128595*.

[2] Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, *10*(1), 115. https://doi.org/10.1186/s40537-023-00792-7

[3] Hradec, J., Ostlaender, N. and Bernini, A., FABLES: Framework for Autonomous Behaviour-rich Language-driven Emotion-enabled Synthetic populations, Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/86682, JRC135070.

[4] The European High Performance Computing Joint Undertaking (EuroHPC JU)

[5] Inondations: l'état de catastrophe naturelle reconnu dans 205 communes du Nord et du Pas-de-Calais https://www.service-public.fr/particuliers/actualites/A16928

[6] Open-Meto data API: https://open-meteo.com/en/docs/historical-weather-api#latitude=50.5296&longitude=2.64&start_date=2023-11-14&end_date=2023-11-16&hourly=temperature_2m,rain

[7] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1-22).

[8] Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., ... & Dai, J. (2023). Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*.

[9] Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X., & Liang, Y. (2023). Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

[10] Ruan, J., Chen, Y., Zhang, B., Xu, Z., Bao, T., Du, G., ... & Zhao, R. (2023). TPTU: large language model-based AI agents for task planning and tool usage. *arXiv preprint arXiv:2308.03427*.

Lead image: Generated by DALL-E3 using the prompt “Synthetic data in policy formulation and adoption”

SUGGESTED CITATION

Hradec, J, Di Leo, M., Kotsev A. (2024). Science for Policy Brief. European Commission - Joint Research Centre, JRC138521.

CONTACT INFORMATION

jrc-data-spaces-kb@ec.europa.eu

EU Science Hub

Joint-research-centre.ec.europa.eu