Clear guidelines needed for synthetic data to ensure transparency, accountability and fairness, study says

April 13 2024

Credit: Pixabay/CC0 Public Domain

Clear guidelines should be established for the generation and processing of synthetic data to ensure transparency, accountability and fairness, a new study says.

Synthetic data—generated through machine learning algorithms from original real-world data—is gaining prominence because it may provide privacy-preserving alternatives to traditional data sources. It can be particularly useful in situations where the actual data is too sensitive to share, too scarce, or of too low quality.

Synthetic data differs from real-world data as it is generated by algorithmic models known as synthetic data generators, such as Generative Adversarial Networks or Bayesian networks.

The study warns existing data protection laws that only apply to personal data are not well-equipped to regulate the processing of all types of synthetic data.

Laws such as the GDPR only apply to the processing of personal data. The GDPR's definition of personal data encompasses 'any information relating to an identified or identifiable natural person'. However, not all synthetic datasets are fully artificial—some may contain personal information or present a risk of re-identification. Fully synthetic datasets are, in principle, exempt from GDPR rules, except when there is a possibility of re-identification.

It remains unclear what level of re-identification risk would be sufficient to trigger their application in the context of fully synthetic data processing. That creates legal uncertainty and practical difficulties for the processing of such datasets.

The study, by Professor Ana Beduschi from the University of Exeter, is published in the journal *Big Data and Society*.

It says there should be clear procedures for calling to account those responsible for the generation and processing of synthetic data. There should be guarantees synthetic data is not generated and used in ways

that bring adverse effects on individuals and society, such as perpetuating existing biases or creating new ones.

Professor Beduschi said, "Clear guidelines for all types of synthetic data should be established. They should prioritize transparency, accountability and fairness. Having such guidelines is especially important as generative AI and advanced language models such as DALL-E 3 and GPT-4—which can both be trained on and generate synthetic data—may facilitate the dissemination of misleading information and have detrimental effects on society. Adhering to these principles could thus help mitigate potential harm and encourage responsible innovation.

"Accordingly, synthetic data should be clearly labeled as such and that information about its generation should be provided to users."

More information: Ana Beduschi, Synthetic data protection: Towards a paradigm change in data regulation?, *Big Data & Society* (2024). DOI: 10.1177/20539517241231277

Provided by University of Exeter