EUROPEAN LAW BLOG
NEWS AND COMMENTS ON EU LAW

ELB Blogpost 39/2022, 22 September 2022
Tags: synthetic data, data synthesis, data protection, GDPR, privacy, law, personal data, non-personal data, pseudonymous data, pseudonymisation, anonymous data, hybrid data, anonymisation, inferences, linkability, distinguishability, singling out, legal aspects, technical aspects, AI, artificial intelligence, machine learning, research
Topics: Data protection and digital governance

On synthetic data: a brief introduction for data protection law for dummies

By César Augusto Fontanillo López and Abdullah Elbi

Synthetic data is attracting increasing attention from technicians and legal scholars in recent years. This is especially noticeable among entities and people working on data-driven technologies, particularly in the artificial intelligence application development and testing sector, where sheer volumes of data are needed. In these circles, synthetic data has become a growing trend under the "fake it till you make it" concept by promising to alleviate existing data access and analytics challenges while respecting data protection rules. Given the rising prospects and acceptance of data synthesis, there is a need to assess the legal implications of its generation and use, the starting point being the legal qualification of synthetic data.

Synthetic data is a broad concept encompassing both personally and non-personally identifiable information. This blog entry focuses, notwithstanding, on the intersection between synthetic data and personal data. The reasons for so doing are that generating synthetic data by means of personal data (including hybrid data) provides a more straightforward assessment and is more suitable for the introductory purposes of this blog entry. At the same time, given the lively academic debate on the concept of personal data, we recognise as particularly relevant to this topic the issues surrounding the qualification as personal data of existing models and background knowledge used as sources for data synthesis. These issues will, however, not be dealt with in this entry.

Thus, the present blog post has been intentionally narrowed down to the delimitation of the notion of synthetic data generated from personal data and to the study of its legal

qualification within the European data protection framework. Three main conclusions are drawn from our analysis: first, full data protection compliance prior to data synthesis would be applicable in many cases; second, according to the identifiability test, synthetic data can be considered pseudonymous or anonymous data depending on the appropriateness of the data synthesis and the related *ex-post* control mechanisms; third, the broader question of legal qualification remains an unresolved issue in light of the discrepancy over the data protection model required by law and doctrine. This blog entry has been heavily influenced by the work of Khaled El Emam.

What is synthetic data?

Synthetic data has many names, such as "fake data" or "artificial data". Regardless of the terminology, synthetic data is, at a fundamental level, data artificially generated from original data that preserves the statistical properties of said original data.

Given an original dataset X, a synthetic dataset X' can be generated by building, for instance, a machine learning model that captures its structure and statistical distribution. The conservation of the statistical properties of X in X' is critical, as it allows data analysts to draw meaningful conclusions from the synthetic data as if they were drawn from the original data.

At the same time, it is possible to induce a certain degree of randomness in the synthetic data generation process, unrelated to the original dataset, to produce data sets with high variability. On this point, it is important to control the level of randomness to ensure that the synthetic data is sufficiently diverse, yet still realistic.

Why synthetic data?

Synthetic data presents technical and legal advantages over original (personal) data processing, which justify its increasing popularity.

From a technical perspective, synthetic data can overcome some of the drawbacks of general database access and use, as well as provide efficient data processing in terms of time and cost. On the first note, a widespread concern about the processing of databases containing original (personal) data relates to the characteristics of such data. Quantitatively, databases may lack a sufficient amount of data. Qualitatively, databases may lack sufficient variability of data. Where database specificities are inadequate to meet particular processing needs, such as training machine learning models or testing mathematical models, their access and use might be hindered or discouraged. In such cases, given the capacity of data synthesis to produce large amounts of data with high variability, this technique can be leveraged to feed models developed by data analysts with valuable data that are both quantitatively and qualitatively tailored to the specific

processing aims, thus overcoming the deficiencies of general database access and use. On the second note, using synthetic data can also ensure a good optimisation of model development processes in terms of time and cost. Under normal circumstances, acquiring and preparing data sets can be a resource-intensive task, especially where data sets need to be labelled for supervised learning. Since the labelling of those data sets is, in many cases, done manually, human and technical resources need to be mobilised, thus making this activity expensive and time-consuming. Against this backdrop, it could be desirable to perform cost-effective preliminary evaluations on synthetic data models to validate assumptions and demonstrate the kind of results that could be obtained with actual models. Data synthesis can provide, in these scenarios, a rapid iteration of model training and experimentation to explore and test hypotheses before engaging in original data collection and processing for actionable model results.

From a legal perspective, synthetic data can respect the right to personal data protection. Following this line, synthetic data is considered a promising alternative to personal data processing by a growing number of scholars. The main argument is that data synthesis could be used as an effective anonymisation technique to access, analyse, share, reuse, and publish data without revealing personal information. Data synthesis is seen, to this extent, as a tool that respects data protection requirements while stimulating technological innovation. The intricacies of considering synthetic data as anonymous data will be discussed below, as part of the legal analysis of this entry.

For the time being, it should be noted that, given the technical and legal advantages of synthetic data, in particular those related to technical efficiency and the alleged respect of data protection rules, the generation and use of synthetic data are increasingly being justified and adopted by a large number of entities and technicians. Some authors have even compared synthetic data to a rich, calorie-free cake because of its advantages over original (personal) data processing.

How to generate appropriate synthetic data?

Data synthesis is subject to a balancing test between utility and anonymity. Utility can be understood as a measure of the satisfaction of synthetic data to produce analysis results similar to those that the original data would produce. Anonymity should be understood as the lack of identifiability as extracted from the definition of personal data in the GDPR. This means that the individual cannot be identified nor be identifiable through direct or indirect identifiers or in combinations with additional pieces of information.

As a rule of thumb, the higher the utility of a synthetic data set, the lower its anonymity. If a synthetic dataset X' maximises utility by fitting the original dataset X very carefully, anonymity would be lost because X' would be a replication of X. If a synthetic dataset X'

maximises anonymity by fitting the original dataset X very carelessly, the utility would be lost because *X'* would be statistically different from *X*. Since, as has been argued elsewhere, it is as relevant to optimise the utility of the synthetic data set as it is to prevent the re-identification of the natural person, the trade-off between utility and anonymity must be, therefore, correctly navigated to generate appropriate synthetic data.

At the same time, one must consider that the very nature of this trade-off is at odds with the plausibility of generating completely anonymous datasets, or datasets with zero risk of re-identification, if utility also needs to be preserved. Consequently, this forces us to consider anonymity in the creation of synthetic datasets in probabilistic terms only. This implies that the determination of whether a synthetic dataset complies with the required anonymity standards or not should be answered, *inter alia*, by considering the probability of re-identification that said synthetic dataset has in relation to an acceptable probabilistic threshold. Following this vein, if data synthesis is carried out poorly, the risk of re-identification can become higher, given the greater chance of record replication. On the contrary, if data synthesis is carried out properly, the risk of re-identification can be minimised. The probability of re-identification can be measured by using different metrics.

Is synthetic data pseudonymous data?

According to the GDPR, pseudonymous data is personal data that cannot be attributed to a specific data subject without the use of additional information, which is to be kept separately and subject to technical and organisational measures to prevent re-identification. While the GDPR does not define the concept of attribution, we understand that the concept refers here to the use of additional information that would make the data subject identifiable. If properly generated, synthetic data cannot be attributed to a specific data subject, given its eugenic nature. This means that the use of additional information may not pinpoint the data subject, therefore circumventing the identifiability test. Nonetheless, synthetic data can still show sufficient structural equivalence with the original dataset or share essential properties or patterns to trigger attribution. For instance, if the synthetic data is generated by one-to-one transformation of the original dataset so that each synthetic datapoint equates to an original data point, source features would be substantially maintained in the synthetic data set and hence would fall under the definition of pseudonymous data. This might be the case where the trade-off of data synthesis is not properly navigated, and the original data set is kept by the controller and used as additional information to draw personal attribution. In such cases, the data protection obligations will apply *tout court.*

Is synthetic data anonymous data?

Anonymous data is defined as information that does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. At its core, anonymisation encompasses not only a set of techniques but also technical and organisational safeguards designed to prevent re-identification over time. Supporters of synthetic data argue that, where synthetic data is properly generated, there is no one-to-one mapping from synthetic records back to the person and therefore consider synthetic data as anonymous data. Of course, such a premise should be considered in statistical terms, taking into account the above-mentioned utility-anonymity trade-off. This means that, where synthetic data is properly generated, it is, statistically speaking, indistinguishable from the original data such as to trigger the anonymisation standard. In these terms, synthetic data is considered to eliminate the risk of re-identification and provide for strong data protection guarantees. Opponents of synthetic data contend that even where it is properly generated, one-to-one relationships are still possible, particularly if the synthetic data set preserves the characteristics of the original data set with high accuracy and/or statistical outliers are present. Based on these assumptions, they consider synthetic data as identifiable information.

Another related concern of synthetic data sets is the possibility of inferring sensitive information about the individual where the identifiability test does not render a positive result. In these cases, the problem of data synthesis amounts to a problem of choice of the desirable regulatory model for data protection law: a model that prevents identifiability and/or a model that prevents information inference. As previously introduced, synthetic data aims to tackle data protection from an identifiability perspective or, in other words, it aims to ensure that the records of the individual would not be singled out or linked. If, however, an adversary knows of the presence of an individual in the original data set, even if that individual cannot be individualised, sensitive inferences might still be possible. According to the opinion of A29WP, to consider personal data as 'truly' anonymous, inferences about the characteristics of the individual must be ruled out. While such a restrictive interpretation of anonymisation enhances the protection of personal data and, consequently, the protection of other fundamental rights and freedoms, it is in material disconnection with data protection law, which focuses on an identifiability substrate, as extracted from the definition of personal data in Article 4(1) GDPR and Recital 26. In other words, while the risk of singling out an individual or disclosing its identity is easily assimilated by the identifiability threshold, the risk of inferring attributes of the person possesses a more difficult accommodation according to the tenor of the law. Following this line, there is a need to discuss the extent to which the recommendations of A29WP help model and enforce the interpretation of the data protection framework and, more generally, the data protection model that society deems adequate.

How to legally qualify synthetic data? Takeaways and open questions

All in all, it should be considered that synthetic data would be, in many cases, generated from original (personal) data, except where data synthesis is based on existing models or background knowledge. While the latter scenarios need separate consideration, where original (personal) data processing, as analysed in this blog entry, is at stake, compliance with the European data protection legislation would be necessary, at least in the phases prior to data synthesis. This implies that the controller would still need to have a lawful basis to collect personal data and be subject to the corresponding data protection obligations in relation to the type and sensitivity of the collected data and the aims pursued. Thereafter, since anonymisation is widely accepted as an instance compatible with the initial purposes, further processing of personal data for data synthesis purposes should not be problematic, provided that the data synthesis is carried out adequately and synthetic data is reliably produced. Only after personal data has been rendered synthetic in such a manner that the data subject is no longer identifiable, European data protection law would be circumvented. Yet one must note that the bar of anonymisation has been set very high by the European legislator. It may comprise anonymisation techniques and post-anonymisation control mechanisms, both technical and organisational. In this sense, the question of whether synthetic data remains anonymous is not a discrete but a continuous issue. It depends on the extent to which the synthetic data deviates sufficiently from the original data to avoid identifiability and the extent to which anonymity is sustained over time. To validate the former, a formal assurance of identifiability must be performed by the controller on the dataset after the data synthesis to validate whether re-identification is possible. To validate the latter, technical measures, such as confidentiality, integrity, availability, and resilience measures, as well as organisational measures, such as security management, incident response, and business continuity, human resources, and test, assessment, and evaluation measures, must be in place. Yet still, the possibility of deducing, with significant probability, attribute values from synthetic datasets remains an unresolved issue.

In a world where trust in anonymisation techniques has eroded in recent years, it is, therefore, necessary to discuss and agree on the desired model on which to build data protection law.

The authors would like to thank Naser Damer, senior researcher at Fraunhofer Institute for Computer Graphics Research IGD, and Khaled El Emam, Canada Research Chair in Medical AI and Professor in the School of Epidemiology and Public Health at the University of Ottawa, for their valuable insights and suggestions in the elaboration of this blog entry.