# On the legal nature of synthetic data

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

The present manuscript attempts to analyze the legal qualification of synthetic data generated from personal data. Three main conclusions are drawn from our legal analysis: first, full data protection compliance prior to data synthesis would be applicable in many cases; second, according to the identifiability test as enshrined in the definition of personal data, synthetic data will be considered pseudonymous or anonymous data depending on the appropriateness of the data synthesis and the related ex-post control mechanisms; third, the question of legal qualification remains, however, an unresolved issue in light of the exegetical discrepancy of the identifiability test and the dissonance over the data protection model required by law and doctrine.

## 1 Disclaimer

Synthetic data is a broad concept encompassing both personally and non-personally identifiable information. This manuscript focuses, notwithstanding, on the intersection between synthetic data and personal data. The reasons for so doing are twofold. First, generating synthetic data by means of personal data (including hybrid data) simplifies our scope of analysis because we avoid entering into the lively academic debate on the concept of personal data prior to the study of the legal qualification of synthetic data. This relaxes the conditions of our assessment as the qualification of existing models and background knowledge used as sources for data synthesis is problematic. Second, the consideration of personal data as a starting point allows us, consequently, to provide a more straightforward assessment of the legal qualification of synthetic data because we can depart from a given premise on which to anchor and study said class, thus considerably reducing the degrees of freedom of our inquiry.

## 2 Introduction

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 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.

In the present manuscript, we attempt to study the legal qualification of synthetic data generated by means of personal data in accordance with the European data protection framework. For this purpose, we focus our analysis on synthetic data as anonymous and pseudonymous data in relation to the identifiability threshold as set out in the General Data Protection Regulation (GDPR). Prior to this analysis, we briefly introduce the concept of synthetic data with the aim of contextualizing the legal debate on identifiability.

Submitted to NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research. Do not distribute.

## 3 Definition

Synthetic data has been named in varying ways, 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. This artificial process of data generation is normally performed by a machine learning model, which captures the structure and statistical distribution of original data to produce a synthetic data. The conservation of the statistical properties of the original data in the synthetic data is crucial, as it allows data analysts to draw meaningful conclusions from the synthetic data as if they were drawn from the original data.

During the synthetic data generation process, or data synthesis, a certain degree of randomness can be induced in the synthetic data, unrelated to the original data, to produce data sets with high variability. This allows the creation of extensive data sets of heterogeneous characteristics that can be used for a multiplicity of purposes. At the same time, the level of randomness in the data generation process can be controlled to ensure that the synthetic data is sufficiently diverse, yet still concordant with the original data.

## 4 Preliminary considerations of the legal nature of synthetic data

The establishment of the legal nature of synthetic data is a complex task comprising the validation of different elements. Some of them are easily assessable, while others are difficult to identify and determine. The ease or difficulty in assessing these elements is conditioned by their degree of interpretability or, in other words, by the level of consistency between the legal qualification made by an individual or entity with respect to the synthetic data and the legal qualification agreed upon by the legal system. If an individual or entity can consistently predict the juridical response of the legal system with respect to the envisaged qualification, the element will be easily validated, and conversely if not.

The degree of interpretability can be influenced by the notion of risk inherent to the European data protection framework and, more particularly, to the concept of personal data, a legal construct for which the risk of re-identification plays a fundamental role in the establishment of its legal nature. Because risk is likely to evolve over time depending on the context, events, time, or agents, the very nature of risk affects the legal determination of synthetic data in the way that, under certain circumstances, synthetic data will be considered anonymous data, while in others, this will not be the case. The consideration of risk is, henceforth, an important element in the analysis of synthetic data, and one that prevents the possibility of legally qualifying synthetic data in an absolute manner.The assessment of the legal nature of synthetic data must be therefore carried out contextually, in attention to the changing nature of risk, so that the legal nature of synthetic data can be sustained or transmuted over time, depending on the circumstances.

## 5 Synthetic data as personal data

Any legal qualification of synthetic data from the perspective of European data protection law must depart from the definition of personal data as it is the dichotomy between personal and non-personal data that synthetic data must navigate. According to Article 4(1) GDPR, personal data means:

> *’any information relating to an identified or identifiable natural person.’*

**A**s previously mentioned, some elements of the definition are easily assessable. These elements are ’any information’ and ’natural person’. In constrast, other elements encounter more complex assessment, such as ’relating to’ and ’identified or identifiable’. The element ’any information’ does not seem to pose major challenges. Indeed, synthetic data can be arguably categorized as information in all circumstances because the very essence of synthetic data is informational, irrespective of its nature, content, or format. This is, regardless of whether synthetic data is accurate or inaccurate, objective or subjective, or generated by certain types of machine learning models or others. One must note here that information seems to be treated by European data protection law as meaning the same as data, *i.e.* it treats both concepts interchangeably. This is an important analytical consideration because, by so doing, the European data protection framework takes an all-encompassing approach

as to what qualifies as information. Conventionally, information differentiates from data in the way that the former is a broader category than the latter. It encompasses data (this is, individual facts or statistics) and the knowledge gained from such data through its analysis and interpretation. By equating information to data, the scope under which personal data is observed becomes irrelevant. In other words, for European data protection law it is arguably of no concern whether personal data can constitute information for certain parties and not for others as long as data is being processed. As a result, the qualification of the element 'any information' constitutes an assessment with a low degree of interpretability.

**M**oreover, synthetic data must refer to the 'natural person' in order to be considered personal data. The natural person is an easily assessable legal construct. It is referred in Article 6 of the Universal Declaration of Human Rights and, in most cases, its definition can be found in Member State's civil law. Conventionally, the natural person refers to the living individual considered as a subject of rights and obligations. Where synthetic data refers to the natural person, it will be considered personal data. For the purposes of our legal analysis, as introduced in the disclaimer section, synthetic data generated from personal data will always concern the natural person within the established constraints, as personal data is the substrate upon which synthetic data is generated. If such was not the case, for instance, because the original data would refer to deceased individuals or to other entities different from the natural person, then such casuistic would generally fall out of the scope of this analysis, as data protection law would not be applicable. One must note, notwithstanding, that this casuistic could still be possible outside the scope of this manuscript.

**A**s explored above, the determination of the former elements constitutes a relatively straightforward undertaking given the high level of foreseeability of their legal validation. In any circumstance, synthetic data will arguably be considered as information and, for the purposes our analysis, synthetic data will always refer to the natural person as a living individual. As regards the natural person, we acknowledge that its validation may be resource-consuming both in terms of time and cost depending on the circumstances. For instance, we conceive it as onerous and tedious the activity of verifying whether the natural persons contained in a data set are alive or not. However, once identified, the determination of this element will most likely not cause interpretative issues.

**A**s opposed to the previous elements, 'relating to' or 'identified or identifiable' are elements for which requires deeper consideration. In the case of 'relating to', its validation merits detailed analysis. On a general note, relatedness can be instantiated if a veridical attribution is established. This is a mental process supported by factual evidence that, despite its eventual mutability over time, can be causally established and verified with respect to a given moment of time. The difficulty in elucidating the relatedness element lies, however, in the angle from which relatedness is established. Since relatedness is a relative concept, or a concept which may vary depending on the entity evaluating such relatedness, the assessment of relatedness can become an uncertain undertaking. For instance, if a given entity possesses certain pieces of information that relate to an individual whose personal data is not held by such entity, it is debatable whether one should consider those pieces of information as relating to said individual or not. The intricacies of this particular consideration will be examined below, while dealing with the element 'identified or identifiable'. In any case, despite this substantive shortcoming, there are also practical considerations that must be tied up with the concept of relatedness. Whereas we acknowledge the difficulty of establishing when information can relate or not to the natural person, in reality, this element will be conditioned, in most cases, by the element 'identified or identifiable'. Simply put, the outcomes of the determination of relatedness would equate, in practical terms, to those obtained in the assessment of the 'identified or identifiable' natural person. The reason for so thinking is that the relative dimension of relatedness is already shared by and considered in the assessment of the 'identified or identifiable' natural person, as will be explained in the following. As long as the 'identified or identifiable' individual is determined, relatedness will follow suit. While such an assumption may entail circular reasoning, value must be placed on it. In fact, taking a different stance would imply epistemological inconsistency in the treatment of the constituent elements of personal data and, consequently, the undermining of the *telos* of European data protection law. We would argue, to this extent, that information about an 'identified or identifiable' individual must always relate to him or her. An instance in which information relates to a non-identified or non-identifiable individual may indeed be possible, but it is at odds with the very idea of consistency. As a matter of fact, information relating to a non-identified or non-identifiable natural person would not merit the consideration of personal data, even if, theoretically, a veridical attribution can be established. Take, for instance, the notions of anonymous or aggregate data. As long as personal data enters into this

domain, it is, in principle, out of the scope of the European data protection framework, irrespective of whether said data could theoretically relate to an individual.

**L**eaving aside the dimension on which to determine the element of relatedness, further accentuation can be put on the ways in which information relates to the individual. According to the doctrine, information can relate to the natural person in content, purpose, or result. For the purposes of our analysis, synthetic data is not likely to relate to the natural person in content because it will most obviously not directly concern the individual as a result of the inclusion of randomness in the data generation process. However, where synthetic data presents data points which preserve the characteristics of the original data with high accuracy and/or statistical outliers are present, synthetic data could relate to the individual in content. Besides this, synthetic data will relate to the natural person more obviously in purpose or result. It will relate to the natural person 'in purpose' where the data controller or a third party makes use of synthetic data with the goal of evaluating, treating in a certain way, or influencing the status or behavior of the data subject. For instance, synthetic data can be used with the aim of evaluating the properties of a certain group without identifying the specific individuals. It will relate to the natural person 'in result' where synthetic data is likely to have an impact on the person's rights and interests. For instance, when the individual is treated differently from other people as a result of the processing of synthetic data. The casuistry of the 'relating to' element can be varied, and contextual elements will need to be taken into account for the legal qualification of synthetic data, thus increasing the complexity of this assessment compared to the previous elements.

**T**he most problematic validation is, however, that concerning the 'identified or identifiable' class, as it constitutes the open yardstick upon which the notion of risk most obviously comes about. At its core, the 'identified or identifiable' element comprises an identification threshold constituted by two limits. The upper limit represents the natural person who has been 'identified', whereas the lower limit represents the natural person who has not been identified yet, but who is possible to be so. A natural person can be 'identified' within the upper limit by reference to direct identifiers, such as the full name or address, or indirect identifiers, such as age, occupation, or place of residence. At the same time, a natural person can also be 'identifiable' within the lower limit by reference to direct or indirect identifiers in combination with other pieces of information, including, but not limited to, gender, ethnic origin, or health status. If the processing of any information relating to the natural person falls within the scope comprised by these two limits, it is presumed that said processing constitutes a risk to the fundamental rights and freedoms of individuals worth triggering European data protection law and, consequently, the corresponding data protection obligations. The difficulty in determining the identifiability threshold is twofold.

From a material point of view, the identifiability threshold is deprived of determinative clarity as to the 'likely reasonably' means to be used to measure the risk of re-identification or, in other words, as to the extent that different means can be considered in assessing whether a person can be singled out or records relating to the individual can be linked to him or her. This problematic has given rise to diametrically opposed lines of thought, better known as the 'absolute' and 'relative' approaches. On the one side, the absolute approach represents the viewpoint under which the measurement of risk is computed by embracing any theoretical probability of identification in relation to the processing of personal data. This approach constitutes the highest level of abstraction in the evaluation of the identifiability threshold. On the other side, the relative approach represents the viewpoint under which the measurement of risk is computed by embracing only the practical probability of identification in relation to the processing of personal data. This approach constitutes the lowest level of abstraction in the evaluation of the identifiability threshold. Both approaches tend to form a dichotomy between theoretical and practical considerations on identifiability. While the absolute approach focuses on what may be possible, the relative approach focuses on what can be probable. For the legal qualification of synthetic data, this would imply the consideration of any possible or probable means, both in the hands of the controller or any third party, that could be used to re-identify the individual. As pointed out previously, we contend that the same framework used for the determination of identifiability should be used for the assessment of relatedness. This implies that, for establishing whether information relates or not to the natural person, the absolute or relative approaches to identifiability should be applied in the same way. Independent assessments and discretionary decisions in the treatment of both elements should be avoided if consistency is to be safeguarded.

**F**rom a formal point of view, the identifiability threshold is deprived of determinative clarity as to the acceptable risk level on which to measure re-identification or, in other words, as to the extent that a data set can be considered to provide anonymity or not. For the legal qualification of synthetic data, this would imply the consideration of the level of identifiability that a dataset should have in order to be considered personal data. This problematic has also given rise to opposed risk assessment frameworks, better known as the 'zero-risk' approach and, in the absence of literature providing a name for it, we would call 'acceptable-risk' approach. On the one side, the zero-risk approach represents the viewpoint under which the risk of re-identification is reduced to zero to consider personal data as anonymous. This approach represents the strictest boundary condition for the evaluation of identifiability. On the other side, the 'acceptable-risk' approach represents the viewpoint under which risk of re-identification is reduced to a certain admissible threshold to consider personal data as anonymous. This approach represents a more relaxed boundary condition for the evaluation of identifiability. Both approaches tend to form, again, a dichotomy between theoretical and practical considerations on identifiability. While the 'zero-risk' approach focuses on what may be desirable for the protection of fundamental rights, the 'acceptable-risk' approach focuses on what can be feasible for the protection of fundamental rights. For the legal qualification of synthetic data, this would imply the reduction of identifiability in a given synthetic dataset to zero (if feasible at all) or to an acceptable level where the risk of re-identification is minimized but not eradicated.

**D**isagreement still exists about the appropriateness of the exposed approaches for the construe of identifiability at the material and formal level. As a result, the determination of the 'identified or identifiable' element and, consequently, of relatedness, still remains an unresolved issue which, in turn, impregnates the assessment of whether synthetic data is personal or not.

## 6 Synthetic data as anonymous data

Anonymous data is defined in Recital 26 GDPR 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.*

**B**ased on this definition, synthetic data is being increasingly defended as an effective anonymisation technique to render personal data anonymous in such a manner that access, analysis, sharing, reuse, and publication of data can be carried out without revealing personal information. One must note, however, that the asseveration of synthetic data as anonymous comes with certain intricacies.

**I**n the first place, data synthesis is subject to a balancing test between utility and anonymity. While 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 in the same terms as described above. As a rule of thumb, the higher the utility of a synthetic data set, the lower its anonymity. If a synthetic data set maximises utility by fitting the original data set very carefully, anonymity would be lost because the synthetic data set would be a replication of the original data set. If a synthetic data set maximises anonymity by fitting the original data set very carelessly, the utility would be lost because the synthetic data set would be statistically different from the original data set. 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.

**A**t the same time, one must consider that the very nature of this trade-off is at odds with the plausibility of generating completely anonymous data sets, or data sets with zero risk of re-identification, if utility also needs to be preserved. This forces one to consider anonymity in the creation of synthetic data sets in probabilistic terms. As a result, the determination of whether a synthetic data set complies with the required anonymity standards or not should be answered, inter alia, by considering the probability of re-identification that said synthetic data set has in relation to an acceptable probabilistic threshold. 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.

**B**ased on the previous assumptions, 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 argued 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. Opponents make use of this argument to consider synthetic data as identifiable information. In contextualising the previous approaches with respect to our analysis on identifiability, one can distinguish that the question of whether synthetic data is considered anonymous or not is actually a problem concerning the formal dimension of the identifiability test or, in other words, a problem of choice between the zero-risk and acceptable-risk approaches, for which we have already noted that no consensus has been reached yet.

**I**n addition to the previous, one must also consider that the application of data synthesis would not circumvent on its own the European data protection framework. At its core, anonymisation encompasses not only a set of techniques, but also technical and organisational safeguards designed to prevent re-identification over time. This is inferrable from the notion of risk inherent to the concept of personal data, which needs to be assessed contextually as argued above, as well as from the tenor of Recital 26 of the GDPR, which emphasises the consideration of the available technology at the time of the processing and technological developments, implying that the future state of the art should be considered while assessing anonymity. According to Recital 26:

> *'[t]o ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.'*

**O**ne final concern of synthetic data is the possibility of inferring sensitive information about the individual yet still where the identifiability test does not render a positive result. This refers to the cases in which the natural person is not identified nor identifiable, but sensitive information can be still inferred from him or her. In technical jargon, this would equate to the risk of attribute disclosure, where one learns something about an individual from the data set with some level of certainty, independent of whether identification concurs or not. 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 Article 29 Working Party (A29WP), an institutional group with authoritative opinion in the field of European data protection law that was replaced by the current European Data Protection Board, 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.

## 7 Synthetic data as pseudonymous data

According to Article 4(5) GDPR, pseudonymous data is:

'*personal data that cannot be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.*'

**W**hile 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 could 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.* One must note, however, that such a conclusion will be determined, in the first place, on the desired approach to identifiability, thus rendering different results based on the chosen data protection model.

**C**losely related to this, it would be possible that, in certain cases, synthetic data might be purposely generated by the data controller with the intention of treating it as pseudonymised data. This might be the case where the utility of the data is considered important, at the expense of anonymity. For these specific contexts, although synthetic data would fall under this category, data synthesis would notwithstanding be considered a useful privacy by design measure for protecting the rights and freedoms of individuals as well as one compliant with the data minimization principle. This can allow researchers, in certain circumstances, to benefit from the advantages of data synthesis while ensuring compliance with data protection law.

# 8 Conclusion

Where data synthesis is carried out from original (personal) data, as analysed in this manuscript, 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. Only after personal data has been rendered synthetic in such a manner that the data subject is no longer identifiable, synthetic data will be considered anonymous. Yet one must note that the bar of anonymisation has been set very high by the European legislator. It may comprise anonymisation techniques, such as data synthesis, 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 the same line, the question of how to categorize synthetic data with respect to the identifiability test and, consequently, to relatedness, poses the challenge of creating societal consensus on which of the approaches, both at material and formal level, are deemed to be more adequate. For this reason, any attempt to legally qualify synthetic data with academic rigour would necessarily need to resolve the prior question of which approaches to follow.