EDPS

16 June 2021

EUROPEAN
DATA
PROTECTION
SUPERVISOR

The EU's independent data protection authority

"Synthetic data: what use cases as a privacy enhancing technology?"

IPEN Webinar on synthetic data

Wojciech Wiewiórowski
European Data Protection Supervisor

Ladies and Gentleman, welcome to our traditional appointment of the annual IPEN workshop, which takes the shape of an online event, again. I am looking forward to meeting you again in person very soon.

Let me refer to the steps that lead to the creation of an original artistic piece.

Regardless of the form of artistic expression, the process seems to follow a, more or less, similar path.

I. In music, for example, this first stage could involve **learning** music notes, tempo, scales and tonality and, later, learning more advanced techniques, such as counterpoint, atonality or use of polyrhythms. This process requires the individual to process a high volume of information and may take hours, days or years.

II. Once the individuals have begun to understand artistic pieces, they may start to **execute**, reproduce the piece.

Both the process of "*understanding the art*" and the process of knowing how to "*reproduce the art*" will greatly improve if artists would gather and correlate influences from several sources and peer executants, all the way to making the artists more complete.

III. Eventually, artists will be able to **create** their own original pieces. While bearing in mind that recreating previous works would result in plagiarism, or tribute (*sometimes it is difficult to draw the line!*) the artist should use the gathered knowledge to produce a relatable, yet different, piece.

Each artist would produce multiple original pieces, all different from each other, while sharing structural similarities that will allow the public to recognise the artist's style, and possibly some of the artist's influences, while still retaining their own identity.

This process is not much different from the process used by artificial intelligence (AI) to create synthetic data.

The process of the so-called data synthesis would take observed data (*training dataset*) and capture its variances in a model from which it would generate new, artificial, machine-made data, with statistical properties similar to the original ones.

Keeping the statistical properties means that anyone analysing the synthetic data, data analysts for example, should be able to draw the same statistical conclusions from the analysis of a given dataset of synthetic data as they would if given the real (original) data.

Although there are several techniques to produce synthetic data, our event today will focus on those based on AI, which are the ones best fitted to deal with high complexity data.

In **AI-based technologies**, the volume of the dataset is much more relevant: the bigger the volume, the more information will be made possible to correlate. The more diversity is provided during the learning stage, the more accurate the system's perception of the environment will be.

Synthetic data is increasingly being used for **machine-learning applications**: for example, a model could be trained on a synthetically- generated dataset with the purpose of transfer learning.

Transfer learning would allow for storing knowledge gained while solving one problem and applying it to a different, but related, one. For instance, knowledge gained while learning to recognise cars could be applied when trying to recognise trucks.

Moreover, we observe an increasing use of synthetic data:

- to test and train systems for fraud or intrusion detection;
- in research fields, to aid in creating a baseline for future studies and testing;
- when organisations want to avoid using personal data to mitigate privacy risks.

Therefore, it comes as no surprise that synthetic data is presented as a sort of **magical solution, almost as a panacea** in some cases, for researchers and practitioners.

It is often boldly claimed that **synthetic data is not real data**.

The underlying rationale being that, synthetic data is the product of an artificial process, not related to real individuals, therefore falling outside of the scope of GDPR.

However, this is not completely clear and not quite pacific: does "real data" stand for "truthful data", maybe?

"Real data" is not a specific notion under the GDPR, which rather discerns the data between personal and non-personal.

The key challenge in synthetic data is whether there is any possible way to establish a correlation between the generated data and the original training dataset that could lead to the identification or singling out of real-life data subjects.

Let me give you a concrete example. The website *thispersondoesnotexist.com* shows one particular use of generative models to produce new data. The system combines pieces from several different pictures of faces to produce a photograph of an inexistent person.

A legitimate question would be: *How would the result look like if the algorithm used pictures of famous individuals in the training dataset? Wouldn't certain facial features be immediately recognisable in that case?*

This and other applications of synthetic data lead us to the key, pivotal, question here**: is synthetic data anonymous data?**

There is a lot of interest these days in synthetic data, and this is not without justification: if synthetic data proves to be truly anonymous, meaning that the information, either does not relate to an identified or identifiable natural person, or was rendered anonymous in such a manner that the individual is not or no longer identifiable, then we are outside the scope of application of the personal data protection framework.

But, if, on the contrary, it is not proven that synthetic data is proper anonymous data, then how should it be qualified under the data protection framework and how where does this fit in?

How to tackle the risk of misuses, if there is any?

Answers to those questions are unclear and still disputed and I expect, and hope, that today's discussion might improve some aspects of our understanding of the subject.

Let me get back to my initial point, the artistic metaphor.

While in music and art a correlation with prior work or influences would still be possible, and most probably natural, the same circumstance in synthetic data, a direct reference to prior input would - as a matter of fact - ***imply the presence of personal data*****, thus defeating the benefits heralded and promised by the use this technology**.

No personal data should result from the data synthesis.

Otherwise, the risk may be that synthetic data may be a “privacy mirage”, as some academics put it, because there is no privacy gains in the end.

A quite specific challenge is the possibility of integrating features from records that stand out within the original dataset into the synthetic data (known in data protection as *outliers*). In extreme cases, this could lead to the inference of the data subject.

Therefore, and similarly to what is done in anonymisation, the process of synthesis should consider a previous preparation of the dataset and, after the model is generated, a privacy assurance assessment should be made to ensure no data from the dataset comes out in the synthetic data.

Society can greatly benefit from AI, but that is not a technological silver bullet and new uses, such as synthetic data, should be carefully accessed.

Accuracy and reliability of the inferences are other aspects to consider, along with the risk of perpetuating bias, which may be inherited by the synthetic data technology as well.

The development and improvement of services and products worldwide is demanding a steadily increasing amount of data, in most cases personal data. The ability to generate, virtually, endless quantities of information, that resemble personal data, seems very appealing.

The EDPS has been closely following the developments of AI. I invite all of you to follow our work, in particular by referring to our Opinion on the European Commission’s White Paper on AI, and the soon to be published Joint EDPS – EDPB Opinion on the Commission’s draft AI Regulation.

Thank you for being with us and for helping us to continue the discussion.