Coursework Task Sheet 4

Evaluation and Communication of Wider issues

CMP5366: Data Management and Machine Learning Operations

Birmingham City University

Kacper Popis – 23161791 – [kacper.popis@mail.bcu.ac.uk](mailto:kacper.popis@mail.bcu.ac.uk)

[Step 1 2](#_Toc625788759)

[Step 2 3](#_Toc1752141065)

[Step 3 4](#_Toc1250862723)

[References 5](#_Toc1399546107)

# Step 1

Data Privacy, in my understanding, is to allow users to limit the amount of data / information that is made publicly accessible to different users / companies over the internet **[1]**. This information can range from data such as the user’s real name, house address or email address.

Data Security, in my understanding, is the process of securing / encrypting user data (as both a business or as an individual) to protect that data from being accessed by unwanted user (like hackers). This process is not limited to only data over the internet, but the actual physical storage and upkeep of user data on things like servers or data lakes / warehouses **[2]**.

Data Ethics, in my understanding, is the concept of both how data is being collected from users, and what the data is being used for. Companies over the years have had to include what data they collect from the user **[3]**, and what is the purpose of the data being collected. An example of this is mobile applications from app stores, they usually contain an EULA that the user has to accept to allow them to use the game. Similar concept applies to marketing purposes, where users can select what sort of data is collected, and how that data is being utilised by the company / partners.

Data Protection Law, in my understanding, is a method of implementing proper steps and methods to protect users and their data from any malicious activity from businesses or individuals. This prevents users from being exploited for their data, and others benefiting from it. For example, one major legislation that has been implement to protect user data is GDPR, which has been enforced in European countries, to protect all European users, no matter if the company is located outside of Europe **[4]**.

# Step 2

What is Synthetic Data and what are the common methods that are currently leveraged to attempt to perform Synthetic Data Generation?

Synthetic Data is data that has been artificially manufactured / deployed as an alternative method to collecting data from real world sources and creating genuine factual data with meaning **[5]**. This sort of data is sometimes used when training AI models to make up for a lack of true data to fabricate accuracy or potential make the AI model more biased towards a specific feature.

It is also a much easier method to “gathering” data as synthetic data can be made digitally on the spot, with minimal effort and minimal costs. Compared to the collection of real world data, this process usually takes much longer to collect the same amount of data that you would generate synthetically **[5]**. Additionally, cost is also a major factor when it comes to collecting data.

Some of the methods used to generate synthetic data are generating data based on statistical distribution or using deep learning models to generate data **[6]**. More and more AI tools have started to become more prevalent in daily life – Chat GPT or Deepseek AI – which an user is able to ask to generate a CSV file based on data inside of an article, publishing or publicly available information across the internet. The model will then collect all the relevant information to the topic presented by the user, and output the data into a CSV file that the user can download and use. Additionally, the user can then ask the AI to refine the data and apply formatting changes to it to make it easier to then reuse the data. Although the data in itself may have been sourced from reliable data outlets, there is some data that will be generated by the model based on pattern recognition of similarities to create fake data to fill in to the user’s request.

# Step 3

One of the legal concerns that are the most prevalent in the solution created would be data security. Although the data is made publicly available on the page of the article, the data stored inside of the machine and the machine learning system itself isn't stored securely. Although the dataset, once it has went through the steps of preprocessing, is stored inside of a database that only specific users are able to get access to, the data in the database isn't encrypted nor are there any proper safety measures implemented to prevent unwanted users from accessing the database. Once an unwanted user gains access to the database they will be able to affect the performance of the AI models as the models function purely based on the data inside of the database.

One of the ways to tighten security of the database would be to (assuming the models have been properly deployed and are publicly available) limit the amount of access users have to both the API and the database in itself – and this can be done by limiting physical access to the system. Additionally, configuring the network’s firewall that the models are hosted on will help protect it from potential DDoS attacks, or limiting the requests that users have to the database from the API, as configuring the API endpoint wrong will allow for users to potentially complete SQL injections to gain access to the database to reduce the performance of the models.

# References

[1] - Cloudflare (2023). What is data privacy? | Privacy definition | Cloudflare UK. Cloudflare. [online] Available at: https://www.cloudflare.com/en-gb/learning/privacy/what-is-data-privacy/. [Accessed. 20th May 2025]

[2] - IBM (2021). Data Security. [online] Ibm.com. Available at: https://www.ibm.com/think/topics/data-security. [Accessed. 20th May 2025]

[3] - Cepelak, C. (2023). An Introduction to Data Ethics: What is the Ethical Use of Data? [online] Datacamp.com. Available at: https://www.datacamp.com/blog/introduction-to-data-ethics. [Accessed. 20th May 2025]

[4] - Wolford, B. (2025). What is GDPR, the EU’s new data protection law? [online] GDPR.eu. Available at: https://gdpr.eu/what-is-gdpr/. [Accessed. 20th May 2025]

[5] - Yasar, K. (2024). What is synthetic data? - Definition from WhatIs.com. [online] SearchCIO. Available at: https://www.techtarget.com/searchcio/definition/synthetic-data. [Accessed. 20th May 2025]

[6] - Turing (2022). Synthetic Data Generation: Definition, Types, Techniques, and Tools. [online] Turing.com. Available at: https://www.turing.com/kb/synthetic-data-generation-techniques#synthetic-data-generation. [Accessed. 20th May 2025]